



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

The Bernstein-von Mises theorem and non-regular models

Citation for published version:

Bochkina, N & Green, PJ 2014, 'The Bernstein-von Mises theorem and non-regular models' *Annals of Statistics*, vol. 42, no. 5, pp. 1850-1878. DOI: 10.1214/14-AOS1239

Digital Object Identifier (DOI):

[10.1214/14-AOS1239](https://doi.org/10.1214/14-AOS1239)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Early version, also known as pre-print

Published In:

Annals of Statistics

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



THE BERNSTEIN–VON MISES THEOREM AND NON-REGULAR MODELS

BY NATALIA BOCHKINA^{†,¶} AND PETER J. GREEN^{§,¶}

*University of Edinburgh[†] and Maxwell Institute[†], University of Bristol[§] and
University of Technology, Sydney[§]*

We study the asymptotic behaviour of the posterior distribution in a broad class of statistical models where the “true” solution occurs on the boundary of the parameter space. We show that in this case the Bayesian inference is consistent, and that the posterior distribution has not only Gaussian components as in the case of regular models (the Bernstein–von Mises theorem) but also has Gamma distribution components that depend on the behaviour of the prior distribution on the boundary and have a faster rate of convergence. We also show a remarkable property of Bayesian inference that for some models, there appears to be no bound on efficiency of estimating the unknown parameter if it is on the boundary of the parameter space. We illustrate the results on a problem from emission tomography.

1. Introduction. The asymptotic behaviour of Bayesian methods has been a long-standing topic of interest, including approximation of the posterior distribution and questions that are important from a frequentist point of view, such as consistency, efficiency and coverage of Bayesian credible regions. For correctly specified regular finite-dimensional models with n independent observations, these properties are captured by the Bernstein–von Mises theorem that implies that the posterior distribution can be approximated in $1/\sqrt{n}$ neighbourhood of the true value of the parameter by a Gaussian distribution with variance given by the Fisher information. [van der Vaart \(1998\)](#) gives a total variation distance version of the theorem, adapted from [Le Cam \(1953\)](#) and [Le Cam and Yang \(1990\)](#). This theorem implies that the prior has no asymptotic influence on the posterior, that posterior inference is consistent and efficient in the frequentist sense, and that posterior credible regions are asymptotically the same as frequentist ones. [Kleijn and van der Vaart \(2012\)](#) has shown that for misspecified regular models, the posterior distribution is also Gaussian in the limit however

[¶]Both authors acknowledge financial support for research visits provided by the EPSRC-funded SuStaIn programme at Bristol University.

Keywords and phrases: approximate posterior, Bayesian inference, Bernstein-von Mises theorem, boundary, non-regular model, SPECT, tomography, total variation distance, variance estimation in mixed models

the credible intervals differ from those constructed using the MLE estimator.

One of the key assumptions of the Bernstein-von Mises (BvM) theorem is that the “true” value of the parameter is an interior point of the parameter space. However, for many problems, including our motivating example of a Poisson inverse problem in tomography, and, more generally for the class of models we consider, this assumption of the BvM theorem does not hold. For the tomography example, the unknown parameter is the vector of image intensities, which are nonnegative and can be zero.

The situation where the unknown parameter can be on the boundary of the parameter support has been addressed in the frequentist literature by studying the asymptotic distribution of the maximum likelihood estimator (Self and Liang 1987; Moran 1971) however it has been studied very little under the Bayesian approach. Dudley and Haughton (2002) studied the asymptotic behaviour of the posterior probability for the unknown parameter to belong to a half-space \mathcal{H} for a regular correctly specified model where they found that if the true value of the parameter belongs to the complement of \mathcal{H} , then the posterior probability of half-space \mathcal{H} goes to zero much faster, namely at least at rate $1/n$ rather than the standard parametric rate $1/\sqrt{n}$ (n here is a sample size), and the upper bound on this posterior probability is exponential. In our setup, this is related to the problem where there is a constraint on the unknown parameter and the unconstrained “true” value of parameter lies outside the constrained set. Also, Erkanli (1994) gave a formula for calculating the expectation of a smooth functional of a 3-dimensional posterior distribution where the unknown parameter is on a smooth boundary.

In this paper, we extend the Bernstein-von Mises theorem by relaxing the assumption that the “true” value of the parameter is interior to the parameter space, in a finite-dimensional setting. We consider a broad class of probability distributions for the data and allow the prior distribution to be zero or infinite on the boundary. We will show that for these models a consequence of relaxing this assumption is twofold: firstly, the rate of convergence is different if the “true” parameter is on the boundary, and secondly, the limit of the posterior distribution has also non-Gaussian components. There are two different types of non-Gaussian components: one is a truncated Gaussian with the same parametric rate of convergence, or its modification if the prior density is not bounded away from zero and infinity on the boundary, and the second one is a Gamma with a faster rate of convergence. An interesting property of the components of the second type is that they do not depend on the data and they are not subject to a bound on efficiency, unlike the “regular” and the first type boundary components.

Under some models with this property, at least part of the data is observed exactly, so perhaps it should not be an unexpected phenomenon (see examples of Poisson and Binomial likelihoods in Section 5). This property is quite remarkable: in principle, it allows the recovery of the unknown parameter on the boundary with an arbitrarily small precision (particularly in the case there is no approximation error), by choosing an appropriate prior distribution, without losing asymptotic efficiency if the parameter is not on the boundary. In frequentist inference, Moran (1971) showed that the asymptotic distribution of a maximum likelihood estimator under the parametric rescaling has a component that is a point mass at the true value of the parameter that is on the boundary. It would be interesting to see if this property remains after the appropriate rescaling.

A related but different problem is a nonregular model where the density of the observations has one or more jumps at a point that depends on the unknown parameter, e.g. $Y_i \sim U[0, \theta]$, $i = 1, \dots, n$, independently. This type of problem has been extensively studied from both frequentist and Bayesian perspective (Ibragimov and Has'minskij 1981; Ghosh *et al.* 1994; Ghosal and Samanta 1995; Ghosal *et al.* 1995; Chernozhukov and Hong 2004; Hirano and Porter 2003). In this case, the rate of convergence of the posterior distribution of the unknown nonregular parameter as a function of n is also faster than the standard parametric rate and it is the same as in the case where the unknown parameter is on the boundary, however there is a crucial difference: in the former case, the posterior distribution has a random bias that depends on the data, whereas in the latter case the posterior distribution asymptotically does not depend on the data.

We motivate our study by presenting in Section 2 an inverse problem from medical imaging; Section 3 establishes the class of models we study. In Section 4 we state the result on the local behaviour of the posterior distribution in a neighbourhood of the limit that is formulated as a modified Bernstein-von Mises theorem, discuss the assumptions, give a heuristic proof and a non-asymptotic version of the result. In Section 5 we illustrate the application of the analogue of the BvM theorem for various examples including the problem of variance estimation in mixed effects models and discuss the choice of the prior distribution. We discuss issues in using the approximation of the posterior distribution in practice and apply it to the data from the motivating example in Section 6. We conclude with a discussion. All proofs are deferred to the Appendix.

Notation. We shall use the default norms $\|z\| = \|z\|_2$ for both vectors and matrices. Define the gradient $\nabla f(\theta)$ of a function f on Θ as a vector of partial derivatives (one-sided if θ is on the boundary of Θ), and $\nabla^2 f(\theta)$ is a

matrix of second derivatives of f (again, one-sided if θ is on the boundary of Θ). We use notation θ_S to define the vector $(\theta_j, j \in S)$ for $S \subset \{1, \dots, p\}$ which also applies to the gradient ∇ , i.e. $\nabla_S f(\theta) = (\nabla_j f(\theta), j \in S)$. We denote a submatrix Σ to subsets S, J by $\Sigma_{S,J} = (\Sigma_{ij}, i \in S, j \in J)$ which also applies to the matrix of second derivatives, i.e. we can write $\nabla_{S,J}^2 f(\theta)$ to denote the corresponding submatrix.

We use $A\mathcal{X} + x_0 = \{Ax + x_0, x \in \mathcal{X}\}$ to denote the image of the affine transformation of set \mathcal{X} given matrix A and vector x_0 . The limit that takes place with \mathbb{P} -probability 1 is denoted by $\mathbb{P}\text{-lim}$, where \mathbb{P} is the true distribution of the data. For $\alpha, a > 0$, $\Gamma(\alpha, a)$ denotes Gamma distribution with density $p(x) = \frac{a^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-ax}$, $x > 0$, $\Gamma(x; \alpha, a)$ is its cumulative distribution function and $\Gamma(dx; \alpha, a)$ is the corresponding probability measure.

2. Motivating example.

2.1. Single photon emission computed tomography. Single photon emission computed tomography (SPECT) is a medical imaging technique in which a radioactively-labelled substance, known to concentrate in the tissue to be imaged, is introduced into the subject. Emitted particles are detected in a device called a gamma camera, forming an array of counts. Tomographic reconstruction is the process of inferring the spatial pattern of concentration of the radioactive isotope in the tissue from these counts.

The Poisson linear model

$$(1) \quad \mathcal{T}Y_i \mid \theta \sim \text{Poisson}(\mathcal{T}A_i\theta), \quad i = 1, \dots, n, \quad \text{independently},$$

is close to reality for the SPECT problem (there are some dead-time effects and other artifacts in recording). Here $\theta = \{\theta_j\}$ represents the spatial distribution of the isotope, typically discretised on a grid, with $\theta_j \geq 0$ for all j , $Y = \{Y_i\}$ the array of the rate of detected photons per time unit, also discretised by the recording process, and \mathcal{T} is the exposure time for photon detection. The array $A = (A_{ij})$ with rows A_i quantifies the emission, transmission, attenuation, decay and recording process; A_{ij} is the mean number of photons recorded at i per unit concentration per unit time at pixel/voxel j , and it is non-negative. Elements of the matrix A are discretised values of the Radon transform, a fact used in some methods of reconstruction.

Since Poisson distributions form an exponential family, this model can be seen as a generalised linear model (Nelder and Wedderburn 1972), with identity link function and dispersion $1/\mathcal{T}$ (see also Example 1 in Section 3.2).

We formalise the notion of small-noise limit for this Poisson model in a practically-relevant way, by supposing that the exposure time for photon detection becomes large, i.e. letting $\mathcal{T} \rightarrow \infty$.

The ‘true image’ θ^* in emission tomography corresponds to a physical reality, the discretised spatial distribution of concentration of a radioactive isotope. Of course, this is non-negative, so we impose the constraints $\theta \in \Theta = [0, \infty)^p \subset \mathbb{R}^p$.

Unless p is too large, i.e. the spatial resolution of θ is too fine, the matrix A is normally of full rank p , and hence the inverse problem is well-posed (although it may be ill-conditioned - see [Johnstone and Silverman \(1990\)](#) for eigenvalues of Radon transform).

See [Green \(1990\)](#) for further detail about this model, and an approach based on EM estimation for MAP reconstruction of θ , in a Bayesian formulation in which spatial smoothness of the solution is promoted by using a pairwise difference Markov random field prior.

2.2. Prior distribution. From the beginning of Bayesian image analysis ([Geman and Geman 1984](#); [Besag 1986](#)), use has been made of Markov random fields as prior distributions for image scenes that express generic, qualitative beliefs about smoothness, yet do not rule out abrupt changes for real discontinuities (for example, at tissue type boundaries in the case of medical imaging).

The prior distribution we consider for the SPECT model is a log cosh pairwise-interaction Markov random field ([Green 1990](#)):

$$(2) \quad p(\theta) \propto \exp \left(-\frac{\delta(1+\delta)}{2\gamma^2} \sum_{j \sim j'} \log \cosh \left(\frac{\theta_j - \theta_{j'}}{\delta} \right) \right), \quad \theta \in \Theta,$$

where $j \sim j'$ stands for j and j' being neighbouring pixels. In this paper the parameters δ and γ are considered to be fixed.

This model has some attractive properties. While giving less penalty to large abrupt changes in θ compared to the Gaussian, it remains log-concave. It bridges the extremes $\delta \rightarrow \infty$, the Gaussian pairwise-interaction prior, and $\delta = 0$, the corresponding Laplace pairwise-interaction model, sometimes called the ‘median prior’.

This distribution is improper since it is invariant to perturbing θ by an arbitrary additive constant, but leads to a proper posterior distribution as long as $\sum_j A_{ij} \neq 0$ for all i .

2.3. Non-standard features of SPECT model. The Bayesian model for SPECT has three non-standard features: (a) the true image θ^* can lie on the boundary of the parameter space $[0, \infty)^p$; (b) if $A_i \theta^* = 0$ for some i , then the distribution of the corresponding Y_i degenerates to a point mass at 0; (c) the prior distribution is not proper.

In the next section we formulate a model that includes the Bayesian SPECT model as a particular case. The approximate behaviour of the posterior distribution of θ for large \mathcal{T} is investigated in Section 6.

3. Model formulation.

3.1. *Likelihood.* We assume that the joint density of the observable responses Y taking values in $\mathcal{Y} \subset \mathbb{R}^n$ (with respect to Lebesgue or counting measure) can be written

$$(3) \quad p_\tau(y|\theta) = C_{y,\tau} \exp \left\{ -\frac{1}{\tau} f_y(\theta) \right\}, \quad y \in \mathcal{Y},$$

for some smooth function $f_y(\theta)$ for $\theta \in \Theta$. We assume that the “true” value of the unknown parameter that generated the data is $\theta^* \in \Theta \subset \mathbb{R}^p$.

Assumption M. We assume that the distribution of Y satisfies the following conditions.

1. For $Y \sim p_\tau(y|\theta^*)$, $\exists f^*(\theta): \Theta \rightarrow \mathbb{R}$ that is a deterministic function independent of τ and such that $f_Y(\theta) \xrightarrow{\mathbb{P}} f^*(\theta)$ as $\tau \rightarrow 0 \quad \forall \theta \in \Theta$.
2. Function $f^*(\theta)$ has a unique minimum over Θ at $\theta = \theta^*$.
3. $\exists \nabla f_Y(\theta^*)$ for \mathbb{P} -almost all $Y \in \mathcal{Y}$, $\exists \nabla f^*(\theta^*)$, and $\mathbb{P}\text{-}\lim_{\tau \rightarrow 0} \nabla f_Y(\theta^*) = \nabla f^*(\theta^*)$.

Further assumptions on $f_y(\theta)$ are given in Section 4.1.

REMARK 1. 1. *Assumption M 1 is satisfied if Y has a distribution from the exponential family in canonical form with dispersion $\tau \rightarrow 0$ (see Example 1 below). In this case, the random variable Y converges in probability to a finite deterministic limit y^* as $\tau \rightarrow 0$, and hence $f_Y(\theta) \rightarrow f_{y^*}(\theta)$ with probability 1. Assumption M 3 also holds in this case. The limit y^* can be interpreted as noise-free data and the dispersion parameter τ is related to the noise level of the observations.*

2. *Assumption M 1 holds for iid observations with $\tau = 1/n$ under mild conditions, with $f^*(\theta) = \mathbb{E}[f_Y(\theta)]$ which equals, up to a constant, the Kullback-Leibler (KL) distance between distributions with densities $p(\cdot | \theta)$ and $p(\cdot | \theta^*)$, and which is minimised over Θ at $\theta = \theta^*$ (see also [Spokoiny \(2012\)](#) in the context of mis-specified models). This assumption holds, for instance, for iid observations from a Cauchy distribution (see Example 2 in the next section). Assumption M 3 is satisfied if the order of taking expectation and differentiation of $f_Y(\theta)$ can be swapped.*

Assumption M 2 means that in the limit $\tau \rightarrow 0$, the most likely value of the unknown parameter under the considered model is its true value θ^* . If Y has a distribution from the exponential family in the canonical form, $f^*(\theta) = f_{y^*}(\theta)$ can be interpreted as the likelihood of the noise-free data, and hence Assumption M 2 means that for the noise-free data, the most likely value of the unknown parameter under the considered model is the true value of the parameter θ^* . See Examples 1 and 2 in the next section.

Assumption M 2 has been used by other authors, for instance, in the context of hidden Markov models by [Douc *et al.* \(2011\)](#) where it was called the *identifiability assumption*, and a finite sample analogue of this assumption in the context of a misspecified model by [Spokoiny \(2012\)](#) (see Example 2).

Assumption M is satisfied by some non-regular models, as the parameter set Θ does not have to be open and the true value of the parameter θ^* can be on the boundary of Θ (see Example 3 in the next section). These assumptions are satisfied for the tomography model discussed in Section 2 where the unknown image θ^* can have zero intensity values in some pixels.

3.2. Examples. Here we show that Assumption M is satisfied for two important classes of models, generalised linear models and iid models, including the case when θ^* is on the boundary of Θ .

EXAMPLE 1. *In the generalised linear models of [Nelder and Wedderburn \(1972\)](#), an important class of nonlinear statistical regression problems, responses y_i , $i = 1, 2, \dots, n$ are drawn independently from a one-parameter exponential family of distributions in canonical form, with density or probability function*

$$p_\tau(y \mid \eta) = \exp \left(\sum_{i=1}^n \left[\frac{y_i b(\eta_i) - c(\eta_i)}{\tau} + d(y_i, \tau) \right] \right),$$

using the mean parameterisation, for appropriate functions b , c and d characterising the particular distribution family. The parameter τ is a common dispersion parameter shared by all responses. Assuming that functions $b(\cdot)$ and $c(\cdot)$ are twice differentiable, the expectation of this distribution is $\mathbb{E}(Y_i) = \eta_i = c'(\eta_i)/b'(\eta_i)$, and the variance is $\text{Var}(Y_i) = \tau[c''(\eta_i)b'(\eta_i) - c'(\eta_i)b''(\eta_i)]/[b'(\eta_i)]^3$. Therefore, Assumption M 1 is satisfied with $f^(\theta) = f_{y^*}(\theta)$ and $y^* = \mathbb{E}Y$ by the Chebyshev inequality as $\tau \rightarrow 0$.*

For $\theta = \eta$, the second assumption is satisfied if θ^ is such that $\nabla f^*(\theta^*) = 0$ since $\theta^* = y^*$ is the solution of*

$$0 = \nabla_i f^*(\theta^*) = -y_i^* b'(\theta_i) + c'(\theta_i) \quad \forall i$$

This is the point of minimum of $f_{y^}(\theta)$ if the Hessian $\nabla^2 f_{y^*}(y^*)$ is diagonal with positive entries. If θ^* is on the boundary, see Example 3 below.*

Therefore, all assumptions are satisfied for this example.

Now consider a generalised linear model with $\eta = A\theta$ and matrix A such that $A^T A$ is of full rank, i.e. so that the likelihood is identifiable with respect to parameter θ . In this case, Assumption M holds with $\theta^* = (A^T A)^{-1} A^T y^*$.

The tomography example given in Section 2 belongs to this class of models, with $\tau = \mathcal{T}^{-1}$, $b(\eta_i) = \log \eta_i$, $c(\eta_i) = \eta_i$, $\eta_i = A_i \theta$ and $\Theta = [0, \infty)^p$.

EXAMPLE 2. Let Y_1, \dots, Y_n be independent identically distributed random variables with unknown parameter $\theta \in \Theta \subseteq \mathbb{R}^p$ with finite p independent of n , with density or probability mass function of Y_i $p(y_i | \theta) = C_{y_i} \exp\{-f_{y_i}(\theta)\}$. Here, $\tau = 1/n$ and $f_y(\theta) = \frac{1}{n} \sum_{i=1}^n f_{y_i}(\theta)$. In this case, $f_{Y_i}(\theta)$ are iid random variables, so, as $n \rightarrow \infty$, Assumption M 1 is satisfied under the conditions of the weak law of large numbers $\forall \theta$ for random variable $f_{Y_i}(\theta)$, which implies that $\exists f^*(\theta): f_Y(\theta) \xrightarrow{\mathbb{P}} f^*(\theta)$ as $n \rightarrow \infty$. If $\mathbb{E}[f_{Y_i}(\theta)] < \infty$ for all $\theta \in \Theta$, then $f^*(\theta) = \mathbb{E}[f_{Y_i}(\theta)]$.

For instance, it is easy to check that Assumption M is satisfied for iid Cauchy random variables Y_i with $f_{Y_i}(\theta) = \log(1 + (Y_i - \theta)^2)$ with $\theta \in \Theta \subseteq \mathbb{R}$ since in this case it is a regular model and the expected score function – the derivative of the expected log likelihood at θ (one-sided if θ is on the boundary) – is finite for all θ and is zero at $\theta = \theta^*$, with positive definite Fisher information matrix.

Now we show that Assumption M 2 is satisfied when θ^* is on the boundary of Θ for some distributions from the exponential family.

EXAMPLE 3. 1. Consider the Poisson distribution: $Y/\tau \sim \text{Pois}(\eta/\tau)$ with $\eta \geq 0$. The log likelihood for η is $\ell(\eta) = [y \log \eta - \eta]/\tau$. If data Y are generated with $\eta = 0$, then we observe $y = 0$ with probability 1, so in this case the likelihood for η is always $-\eta/\tau$ which is maximised over $\eta \geq 0$ at $\eta = 0$, i.e. the true value of η .

2. For Binomial distribution $Y \sim \text{Bin}(n, \eta)$, the log likelihood for $\eta \in [0, 1]$ is $\ell(\eta) = y \log(\eta) + (n - y) \log(1 - \eta)$. If the true value of η is 1, then $\mathbb{P}(Y = 1) = n$ and the likelihood for η is $\ell(\eta) = n \log(\eta)$, which is maximised over $[0, 1]$ at $\eta = 1$, i.e. again we recover the true value, so Assumption M 2 is satisfied for this model.

The same holds for the other boundary point $\eta = 0$, and also for multinomial and Negative Binomial distributions.

3.3. Bayesian formulation. We adopt a Bayesian paradigm, using a σ -finite prior measure $\pi(d\theta)$ on Θ . Thus the posterior distribution satisfies

$$(4) \quad \pi(d\theta|y) \propto \exp(-f_y(\theta)/\tau) \pi(d\theta), \quad \theta \in \Theta.$$

Here we do not assume that the prior distribution is proper, nor do we assume that it is bounded away from 0 and infinity on the boundary of Θ (see Assumption P in Section 4.1).

4. The analogue of the Bernstein-von Mises theorem. The limiting statements are given in terms of $\sigma = \sqrt{\tau}$.

4.1. Notation and assumptions. Define

$$(5) \quad \begin{aligned} S &= \{j \in 1, 2, \dots, p : \nabla_j f^*(\theta^*) \neq 0\}, \\ S^* &= \{j \in 1, 2, \dots, p : j \notin S \text{ \& } \theta_j^* = 0\}. \end{aligned}$$

These two sets will determine the directions where the standard Bernstein-von Mises theorem does not hold. We will also need $\bar{S} = \{1, 2, \dots, p\} \setminus S$. Note that $S^* \subseteq \bar{S}$.

We define

$$(6) \quad p_0 \stackrel{\text{def}}{=} |\bar{S}|, \quad p_1 \stackrel{\text{def}}{=} |S|, \quad p_0^* \stackrel{\text{def}}{=} |\bar{S}| - |S^*|;$$

note that $p_0 + p_1 = p$. We then introduce a permutation of coordinates of θ , defined by a matrix V , so that V maps the first $(p_0 - p_0^*)$ coordinates to $\bar{S} \setminus S^*$, the next p_0^* to S^* , and the last p_1 to S . The first p_0 rows of V will be denoted V_0 and the remainder V_1 .

To describe the limit of the posterior distribution, we will need to introduce a modification of a Gaussian distribution truncated to $\mathcal{V}_0 = \mathbb{R}^{p_0 - p_0^*} \times \mathbb{R}_+^{p_0^*}$, with the corresponding measure of any measurable $\mathcal{B} \subset \mathcal{V}_0$ defined by

$$(7) \quad \Phi_{p_0^*}(\mathcal{B}; a_0, \Omega_{00}, \alpha_0) = \frac{\int_{\mathcal{B}} \prod_{i=1}^{p_0^*} x_{p_0 - p_0^* + i}^{\alpha_{0,i} - 1} e^{-(x - a_0)^T \Omega_{00} (x - a_0)/2} dx}{\int_{\mathcal{V}_0} \prod_{i=1}^{p_0^*} x_{p_0 - p_0^* + i}^{\alpha_{0,i} - 1} e^{-(x - a_0)^T \Omega_{00} (x - a_0)/2} dx},$$

where $a_0 \in \mathbb{R}^{p_0}$, $\alpha_0 \in (0, \infty)^{p_0^*}$, Ω_{00} - $p_0 \times p_0$ positive definite matrix. Note that this distribution is Gaussian if $p_0^* = 0$, and it is a truncated Gaussian if $\alpha_{0,i} = 1$ for all i .

In addition to Assumption M (Section 3.1) with $\tau = \sigma^2$, we make five assumptions: about the boundary of set Θ (Assumption B), that the derivatives of the log-likelihood $f_Y(\theta)$ are uniformly bounded for θ close to θ^* with

high probability (Assumption S), about continuity of the log likelihood and its derivatives with respect to data Y (Assumption C), that the posterior distribution is proper and how the prior density behaves on the boundary of Θ (Assumption P), and that the posterior distribution is concentrated in a neighbourhood of θ^* with high probability (Assumption L).

Local neighbourhoods. First we define local neighbourhoods of θ^* :

$$(8) \quad B_\delta(\theta^*) = \{\theta \in \Theta : \theta - \theta^* \in B_{2,p_0}(0, \delta_0) \times B_{\infty,p_1}(0, \delta_1)\},$$

where $\delta = (\delta_0, \delta_1)$, $\delta_0, \delta_1 > 0$ and $B_{q,s}(z_0, r) = \{z \in \mathbb{R}^s : \|z - z_0\|_q < r\}$.

Assumption B (on boundary of Θ).

1. $\Theta \subseteq [0, \infty)^p$.

2. $\exists c_0, c_1 > 0: \{v \in B_{2,p_0}(0, c_0) : v_{(p_0-p_0^*+1):p_0} \geq 0\} \times [0, c_1]^{p_1} \subseteq V(\Theta - \theta^*)$.

Assumption S (smoothness in θ).

$\exists \delta_0, \delta_1 > 0$ depending on σ such that

1) $\delta_0 \rightarrow 0, \delta_1 \rightarrow 0, \delta_0/\sigma \rightarrow \infty, \delta_1/\sigma^2 \rightarrow \infty$ as $\sigma \rightarrow 0$,

2) $\exists \nabla f_Y(\theta), \nabla_{\bar{S}, \bar{S}}^2 f_Y(\theta)$ \mathbb{P} -almost everywhere on $B_\delta(\theta^*)$,

3)

$$\mathbb{P}\text{-}\lim_{\sigma \rightarrow 0} \sup_{\theta \in B_\delta(\theta^*)} |\nabla_j f_{Y(\omega)}(\theta) - \nabla_j f_{Y(\omega)}(\theta^*)| = 0 \quad \forall j \in S,$$

$$(9) \quad \mathbb{P}\text{-}\lim_{\sigma \rightarrow 0} \sup_{\theta \in B_\delta(\theta^*)} |\nabla_{ij}^2 f_{Y(\omega)}(\theta) - \nabla_{ij}^2 f_{Y(\omega)}(\theta^*)| = 0 \quad \forall i, j \in \bar{S}.$$

Assumption C (continuity in Y).

1. $\exists p_0 \times p_0$ positive definite matrix Ω_{00} such that

$$(10) \quad \mathbb{P}\text{-}\lim_{\sigma \rightarrow 0} \nabla_{\bar{S}, \bar{S}}^2 f_{Y(\omega)}(\theta^*) = \Omega_{00}.$$

2. $\|\sigma^{-1} \nabla_{\bar{S}} f_{Y(\omega)}(\theta^*)\| < \infty$ \mathbb{P} -almost everywhere for small enough σ .

Assumption P (on prior distribution).

The σ -finite measure $\pi(d\theta)$ on Θ satisfies the following conditions.

1. $\exists \sigma_0 > 0: \forall \sigma \leq \sigma_0, \int_{\Theta} e^{-f_Y(\theta)/\sigma^2} \pi(d\theta) < \infty$ for \mathbb{P} -almost all $y \in \mathcal{Y}$.

2. For $\theta \in B_\delta(\theta^*)$, $\exists p(\theta) \geq 0: \pi(d\theta) = p(\theta)d\theta$.

3. $\exists C_\pi \in (0, \infty)$ and $\exists \alpha_j \in (0, \infty)$ for $j \in S \cup S^*$ independent of σ , $\exists \Delta_\pi = \Delta_\pi(\delta) \geq 0$, such that $\Delta_\pi \rightarrow 0$ as $\sigma \rightarrow 0$ and for $\theta \in B_\delta(\theta^*)$,

$$C_\pi(1 - \Delta_\pi) \leq p(\theta) \times \prod_{j \in S \cup S^*} \theta_j^{-(\alpha_j - 1)} \leq C_\pi(1 + \Delta_\pi).$$

Denote $\alpha_0 = \alpha_{S^*}$, $\alpha_1 = \alpha_S$.

Assumption L.

Assume $\mathbb{P}(\Delta_0(\delta) \rightarrow 0) \rightarrow 1$ as $\sigma \rightarrow 0$, where
(11)

$$\Delta_0(\delta) = \sigma^{-p_0 - \sum_{j=1}^{p_0^*} (\alpha_{0,j} - 1) - 2 \sum_{j=1}^{p_1} \alpha_{1,j}} \int_{\Theta \setminus B_\delta(\theta^*)} e^{(f_Y(\theta^*) - f_Y(\theta))/\sigma^2} \pi(d\theta).$$

Now we discuss the assumptions.

REMARK 2 (Assumption B). *Under Assumption B, the complement of the polar cone to set Θ at point θ^* coincides with Θ in a small enough neighbourhood of θ^* ; this significantly simplifies the analytic argument. This property holds for other polyhedral boundaries, since these assumptions hold for $A\theta$ if they hold for θ . In this way, Assumption B on the parameter set Θ that only zero values θ^* correspond to the boundary points, can be relaxed to any polyhedral boundary. For a set Θ that does not satisfy these conditions, the support of the posterior distribution in the limit may depend on the complement of the polar cone to set Θ at point θ^* (see also [Shapiro \(2000\)](#)).*

REMARK 3 (Assumption P). *We assume that the posterior distribution is proper but we do not assume that the prior measure itself is proper, nor do we assume that the prior density is finite and bounded away from 0 on the boundary of the parameter space. If $\alpha_j = 1$ for all j , this corresponds to the case of a locally flat prior that is finite and bounded away from 0 in a neighbourhood of θ^* . In particular, the log cosh Markov random field prior distribution that was discussed in Section 2 for the motivating example, satisfies these conditions with $\alpha_j = 1 \forall j \in S \cup S^*$. Other improper priors such as the Jeffreys prior for a Poisson likelihood, as well as the conjugate Gamma prior and Beta prior conjugate to a Binomial likelihood, satisfy this assumption (see examples in Section 5). Assumption $\Delta_\pi \rightarrow 0$ as $\sigma \rightarrow 0$ implies that function $p(\theta) \prod_{j \in S \cup S^*} \theta_j^{1-\alpha_j}$ is continuous at θ^* .*

A simple rule to verify Assumption L is presented in Lemma 3 (Section 5).

4.2. *The main result.* Before presenting the main result, we state two preliminary lemmas. Firstly, we show that the elements θ_S^* are on the boundary of Θ and secondly, we study properties of the derivatives of $f^*(\theta)$.

LEMMA 1. *If Assumption M in Section 3.1 and Assumption B in Section 4.1 hold, then $\theta_S^* = 0$ and vector $\nabla_S f^*(\theta^*)$ has positive coordinates.*

If also $\nabla_{\bar{S}, \bar{S}}^2 f_Y(\theta^) \xrightarrow{\mathbb{P}} \nabla_{\bar{S}, \bar{S}}^2 f^*(\theta^*)$ as $\sigma \rightarrow 0$, then matrix $\Omega_{00} = \nabla_{\bar{S}, \bar{S}}^2 f^*(\theta^*)$ is positive semi-definite.*

Define the following scaling transform $\mathcal{S} = \mathcal{S}_\sigma: \Theta - \theta^\star \rightarrow \mathbb{R}^{p_0} \times \mathbb{R}_+^{p_1}$:

$$(12) \quad \mathcal{S}(\theta - \theta^\star) = D_\sigma^{-1} V(\theta - \theta^\star),$$

where $D_\sigma = \text{diag}(\sigma I_{p_0}, \sigma^2 I_{p_1})$ and $V = (V_0^T : V_1^T)^T$ is defined in Section 4.1. This corresponds to rescaling each of the two subsets of coordinates differently, namely considering $(\theta_{\bar{S}} - \theta_{\bar{S}}^\star)/\sigma$ and $(\theta_S - \theta_S^\star)/\sigma^2$. In the next lemma we study the image of this transformation in the limit.

LEMMA 2. *Let Assumption B in Section 4.1 hold, with $\delta_0, \delta_1 \rightarrow 0$ and $\delta_0/\sigma \rightarrow \infty$ and $\delta_1/\sigma^2 \rightarrow \infty$ as $\sigma \rightarrow 0$. Then,*

$$\limsup_{\sigma \rightarrow 0} \mathcal{S}_\sigma(\Theta - \theta^\star) = \mathbb{R}^{p_0 - p_0^\star} \times \mathbb{R}_+^{p_0^\star + p_1}.$$

The proofs of both lemmas are given in Appendix A.1.

The limit of the posterior distribution is described by the following parameters: $\alpha_0 = \alpha_{S^\star}$, $\alpha_1 = \alpha_S$, and

$$(13) \quad a_1 = \mathbb{P}\text{-}\lim_{\sigma \rightarrow 0} [\nabla_S f_Y(\theta^\star)], \quad a_0(\omega) = -\sigma^{-1} \Omega_{00}^{-1} \nabla_{\bar{S}} f_{Y(\omega)}(\theta^\star),$$

$$\Omega_{00} = \mathbb{P}\text{-}\lim_{\sigma \rightarrow 0} [\nabla_{\bar{S}, \bar{S}}^2 f_Y(\theta^\star)].$$

Existence of the limits follows from Assumptions C and M, which also imply that vector a_1 has positive coordinates (Lemma 1). The matrix $\sigma^{-2} \Omega_{00}$ is an analogue of the Fisher information for parameter $\theta_{\bar{S}}$.

In the theorem below, which is an analogue of the Bernstein–von Mises theorem, we show that the posterior distribution of $\mathcal{S}(\theta - \theta^\star)$ converges to a finite limit.

THEOREM 1. *Consider the Bayesian model defined in Section 3 under Assumption M and such that Assumptions B, P, S, C and L stated in Section 4.1 hold.*

Define a random probability measure on $\mathcal{V}_0 \times \mathbb{R}_+^{p_1}$, with $v = (v_0, v_1)$:

$$\mu^\star(\omega)(dv) = \Phi_{p_0^\star}(dv_0; a_0, \Omega_{00}, \alpha_0) \times \Gamma_{p_1}(dv_1; \alpha_1, a_1),$$

where $\mathcal{V}_0 = \mathbb{R}^{p_0 - p_0^\star} \times \mathbb{R}_+^{p_0^\star}$, $\Phi_{p_0^\star}(dv_0; a_0, \Omega_{00}, \alpha_0)$ is the modified Gaussian distribution defined by (7), and $\Gamma_{p_1}(\cdot; \alpha_1, a_1)$ is the probability measure of a p_1 -dimensional vector ξ with independent coordinates $\xi_i \sim \Gamma(\alpha_{1,i}, a_{1,i})$.

Then, with transform \mathcal{S} defined by (12), as $\sigma \rightarrow 0$,

$$\|\mathbb{P}_{\mathcal{S}(\theta - \theta^\star)|Y} - \mu^\star\|_{TV} \xrightarrow{\mathbb{P}} 0.$$

The proof is given in Appendix A.2. If θ^* is an interior point, then $p_1 = p_0^* = 0$, the corresponding factor in the definition of μ^* disappears, and the limit is a Gaussian distribution, i.e. this becomes the classical Bernstein-von Mises theorem. If $\theta^* = 0$ and $\alpha_j = 1$ for all $j \in S^*$, then the limit of the posterior distribution is a product of Gamma and truncated Gaussian distributions.

Note that the parameter on the boundary can exhibit two types of limiting behaviour. If the model is “regular” for parameter θ_j and $\theta_j^* = 0$ (i.e. $j \in S^*$), then the rate of convergence is still σ^{-1} and the limit of the rescaled posterior is Gaussian, possibly modified by non-regular behaviour of the prior density on the boundary. If the prior density of θ_j behaves locally as a constant around θ^* , then the limit is a truncated Gaussian distribution, and this parameter can be dependent on other “regular” parameters in the limit. However, if the model is “non-regular” for parameter θ_j ($j \in S$), then the situation changes: the rate of convergence is now faster (σ^{-2} instead of σ^{-1}), it is independent of other parameters and the limiting distribution is Gamma. See examples in Section 5.

REMARK 4. *The key property of the posterior distribution when the true parameter is on the boundary, is that the gradient of the log likelihood at this point does not asymptotically vanish, so that in some directions a leading term at the Taylor expansion of log posterior density is linear rather than quadratic, as in the case when θ^* is an interior point. If the prior density at θ^* is bounded away from 0 and infinity, then the limit of the posterior in these directions is an exponential distribution; if the prior density has an additional polynomial term at a neighbourhood of $\theta_j^* = 0$ then the limit is a Gamma distribution.*

If the prior distribution behaves like a positive constant on the boundary or the “regular” part of the parameter is not on the boundary, then the limiting distribution $\mu^*(\omega)$ has a simple form.

COROLLARY 1. *Assume that Assumption P is satisfied with $\alpha_j = 1$ for $j \in S^*$, or the set S^* is empty. Then, under the conditions of Theorem 1, the limiting measure $\mu^*(\omega)$ is defined by*

$$\mu^*(\omega) = [\mathcal{N}_{p_0}(a_0(\omega), \Omega_{00}^{-1}) \mathbf{1}_{\mathcal{V}_0}] \times \otimes_{i=1}^{p_1} \Gamma(\alpha_{1,i}, a_{1,i}),$$

where $\mathbf{1}_{\mathcal{V}_0}$ is the indicator function of set \mathcal{V}_0 , and $\mathcal{N}_{p_0}(a_0(\omega), \Omega_{00}^{-1}) \mathbf{1}_{\mathcal{V}_0}$ denotes the normal distribution truncated to \mathcal{V}_0 and normalised to be a probability measure.

In particular, in the prior distribution behaves as a constant in a neighbourhood of θ^ ($\alpha_{1,j} = 1$ for all j), then the limit of θ_S/σ^2 is multivariate exponential.*

Therefore, if the set S^* is empty or $\alpha_j = 1$ for $j \in S^*$, Bayesian inference for the “regular” parameter $\theta_{\bar{S}}$ that is an interior point of the parameter set, is asymptotically equivalent to the frequentist inference, which is one-sided for θ_{S^*} . Under these conditions, asymptotic efficiency of posterior inference for $\theta_{\bar{S}}$ is not affected by the presence of “non-regular” parameters θ_S .

4.3. Inference on the boundary. We can see that for $\theta_{\bar{S}}$ the standard Bernstein–von Mises theorem holds, i.e. inference for $\theta_{\bar{S}}$ is asymptotically equivalent to the efficient frequentist inference, under the assumption that the prior density in the neighbourhood of $\theta_{\bar{S}}$ is bounded away from 0 and infinity, a standard assumption of the BvM theorem. However, inference for θ_S is different. The first key difference is that there is no need to require a similar assumption on the prior distribution: even if the prior density tends to infinity or is zero (both of a polynomial order) on the boundary, for the iid observations with $\sigma^2 = 1/n$, Bayesian inference is still consistent, at a rate faster than the parametric \sqrt{n} rate. The second difference is that the limit of the rescaled and recentred posterior distribution for θ_S^* does not depend on the observed data. These two properties lead to the third important difference which is the formulation of efficiency of the estimation procedure for these “non-regular” parameters. The latter point is elaborated below.

If the prior density is not bounded away from 0 and infinity at θ^* , the limit of the posterior distribution depends on the behaviour of the prior distribution on the boundary via exponents α_j , $j \in S \cup S^*$. These exponents are a construct of the statistician and do not depend on the underlying data or its model and can be chosen arbitrarily. If $\alpha_j > 1$ then the prior density is 0 at the true value θ^* , and if $\alpha_j < 1$, the prior density of θ_j tends to infinity as $\theta_j \rightarrow \theta_j^*$. The length of the asymptotic posterior credible interval for θ_j decreases to 0 as $\alpha_j \rightarrow 0$, hence it is possible to recover the true value on the boundary as precisely as desired, possibly up to approximation error, without affecting the coverage of the posterior credible regions for $\theta_{\bar{S}}$ asymptotically. This can be done with the prior distribution independent of θ^* (i.e. without knowledge of S), by choosing a prior density so that $\lim_{\theta \rightarrow 0, \theta_j \geq 0 \forall j} p(\theta) \prod_{j=1}^p \theta_j^{1-\alpha_j} = \text{const} \in (0, \infty)$ for a small positive α . Therefore, *it is advantageous to have a prior density that tends to infinity on the boundary of the parameter space*, as long as the posterior distribution – and its limit $\mu^*(\omega)$ – remains proper. This property raises the question about the formulation of efficiency in this case, as, from the theoretical per-

spective, there appears to be no lower bound on the length of the credible interval as in the regular case.

This property supports the use of Jeffreys priors for Poisson likelihoods and for the probability of success in Bernoulli trials in the absence of subjective information, since it leads to smaller credible sets than using a uniform prior (Example 5 in Section 5).

Now let us see what happens if the parameter on the boundary turns out to be “regular”, i.e. $\nabla_j f^*(\theta^*) = 0$ and $\theta_j^* = 0$ ($j \in S^* \subset \bar{S}$). The posterior distribution of the rescaled and recentred parameter is approximately modified Gaussian and may be correlated with other parameters. Thus, if this parameter is estimated with a higher precision, this will imply that there will be a loss in efficiency in estimating this and any other correlated parameters. Therefore, in this case it may best to choose a prior whose density is bounded away from zero and infinity everywhere on Θ ($\alpha_j = 1 \ \forall j$). This point is illustrated on a mixed model in Section 5.

We shall see in Section 5 that in a number of models the parameter on the boundary can be either only regular or only non-regular. However, in the motivating SPECT example, both types of boundary behaviour can occur, hence the chosen prior, that satisfies condition P with $\alpha_j = 1$ for all $j \in S \cup S^*$, results in asymptotically efficient inference with respect to the regular parameters.

4.4. Non-asymptotic upper bound. We will also state a nonasymptotic bound on the distance between the posterior distribution of the rescaled parameter and its limit.

PROPOSITION 1. *Let the assumptions of Theorem 1 hold, and define the following events:*

$$\begin{aligned} \mathcal{A}_0 &= \{\omega : \|\nabla_{\bar{S}, \bar{S}}^2 f_{Y(\omega)}(\theta^*) - \Omega_{00}\| \leq M_0 \delta_0\}, \\ \mathcal{A}_1 &= \{\omega : \|[\nabla_S f_{Y(\omega)}(\theta^*) - a_1]\|_\infty \leq M_1 \delta_1\}, \\ \mathcal{A}_2 &= \{\omega : \|\nabla_S f_{Y(\omega)}(\theta^*) - \nabla_S f_{Y(\omega)}(\theta)\|_\infty \leq M_0 \delta_1 \ \forall \theta \in B_\delta(\theta^*)\}, \\ \mathcal{A}_3 &= \{\omega : \|\nabla_{\bar{S}, \bar{S}}^2 f_{Y(\omega)}(\theta^*) - \nabla_{\bar{S}, \bar{S}}^2 f_{Y(\omega)}(\theta)\| \leq M_1 \delta_0\} \ \forall \theta \in B_\delta(\theta^*), \\ \mathcal{A}_4 &= \{\omega : \|\nabla_{\bar{S}} f_{Y(\omega)}(\theta^*)\| \leq \delta_0/2\}. \end{aligned}$$

Assume that $\delta_0, \delta_1 > 0$ satisfy the following conditions

$$\delta_1 < a_{\min}/(2M_1), \ \delta_0 < \lambda_{\min}(\Omega_{00})/(2M_0), \ \delta_0 \leq \|\theta_{\bar{S}}^*\|, \ \delta_0 \leq c_0, \ \delta_1 \leq c_1,$$

where $a_{\min} = \min_j a_{1,j}$, $\lambda_{\min}(\Omega_{00})$ is the smallest eigenvalue of Ω_{00} , and c_0, c_1 are constants from Assumption B.

Then, on $\mathcal{A} = \mathcal{A}_0 \cap \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 \cap \mathcal{A}_4$,

$$\begin{aligned} \|\mathbb{P}_{S(\theta-\theta^*)|Y} - \mu^*\|_{TV} &\leq 2 \max \left\{ C_0 \delta_0, C_{\alpha_0} \left[1 - \Gamma \left(\frac{\lambda_{\min}(\Omega_{00}) \delta_0^2}{8\sigma^2}; \frac{p_{\alpha_0}}{2}, 1 \right) \right] \right\} \\ &+ 2 \max \left\{ C_1 \delta_1, p_1 \left[1 - \min_j \Gamma \left(\frac{a_{1,j} \delta_1}{\sigma^2}; \alpha_{1,j}, 1 \right) \right] \right\} \\ &+ C_2 \Delta_\pi + C_\Delta \Delta_0(\delta), \end{aligned}$$

where $p_{\alpha_0} = p_0 + \sum_{j=1}^{p_0^*} (\alpha_{0j} - 1)$, $\lambda_{\min}(\Omega_{00})$ is the smallest eigenvalue of Ω_{00} and the constants are defined in the proof (Appendix A.2).

The proof is given in Appendix A.2. The upper bound implies that for the total variation to be small in practical applications, the dimensions p_k should not be too large compared to the corresponding rate, the smallest eigenvalue of the precision matrix Ω_{00} cannot be too small, namely that $\lambda_{\min}(\Omega_{00}) \delta_0^2 / \sigma^2$ should be large, and that the combination of parameters $(\alpha_{1,j}, a_{1,j})$ should be such that value δ_1 / σ^2 is far in the tail of all corresponding Gamma distributions. If $\alpha_{1,j} = 1$ for all j , this implies that the smallest value a_{\min} of the parameter a_1 should not be too small, i.e. $a_{\min} \delta_1 / \sigma^2$ should be large.

It is interesting to note that the values of δ_k minimising the local upper bound (the first two lines of the upper bound) coincide with an upper bound on the Ky Fan distance between the posterior distribution and its limit, point mass at θ^* , on the corresponding subset of the parameter space which are $\delta_0 = C_{\Omega_{00}} \sigma \sqrt{\log(1/\sigma)}$ and $\delta_1 = C_{a_1} \sigma^2 \log(1/\sigma)$ (Bochkina 2012).

5. Examples. We start with a rule to verify Assumption L.

LEMMA 3. Take $\delta_0, \delta_1 > 0$ such that $\delta_0, \delta_1 \rightarrow 0$, and assume that

$$\sup_{\Theta \setminus B_\delta(\theta^*)} [f_Y(\theta^*) - f_Y(\theta)] \leq -C_{\delta_0} \sum_{j \in \bar{S}} |\theta_j - \theta_j^*| - C_{\delta_1} \sum_{j \in S} |\theta_j - \theta_j^*|$$

for some $C_{\delta_0}, C_{\delta_1} > 0$ with probability close to 1 for small enough σ , and that $\exists \alpha_j > 0$, $j = 1, \dots, p$, and $C_{\pi_0} > 0$ such that $\forall \theta \in \Theta$,

$$\frac{\pi(d\theta)}{d\theta} \leq C_{\pi_0} \prod_{j \in \bar{S}: |\theta_j| < \delta_0 / \sqrt{p_0}} \theta_j^{\alpha_j - 1} \prod_{j \in S: \theta_j < \delta_1} \theta_j^{\alpha_j - 1}.$$

Then, if $C_{\delta_0} \delta_0 / \sigma^2 \rightarrow \infty$ and $C_{\delta_1} \delta_1 / \sigma^2 \rightarrow \infty$, then $\Delta_0(\delta) \rightarrow 0$ as $\sigma \rightarrow 0$ with probability 1, i.e. Assumption L is satisfied.

The proof is given in Appendix A.3. Here are examples where the posterior distribution has non-Gaussian components.

EXAMPLE 4 (Poisson likelihood 1). Consider $Y = (Y_1, Y_2)^T$ where $Y_1/\tau \sim \text{Pois}(\theta_1/\tau)$ and $Y_2/\tau \sim \text{Pois}((\theta_2 + a\theta_1)/\tau)$ independently, $\theta_1, \theta_2 \geq 0$, for some known $a > 0$. This model is identifiable. Assume the prior is uniform on the support of $\theta = (\theta_1, \theta_2)^T$ which is improper.

Assumptions of Lemma 3 are satisfied for independent Poisson random variables with

$$C_{\delta 0} = 0.5\delta(\delta + \sqrt{p_0}y_{\min}^*)^{-1}y_{\min}^*, \quad C_{\delta 1} = \min_j(a_{1,j})$$

where $y_{\min}^* = \min_{j: y_j^* > 0} y_j^*$ due to inequality $\log(1+x) - x \leq -\frac{a}{a+1}x$ for $x > a > 0$, for small enough σ . Hence, all assumptions of Theorem 1 are satisfied for this model.

Let the true value of θ that generated this be $\theta^* = (0, 1)^T$. Then, $P(Y_1 = 0) = 1$ and the likelihood is non-regular.

For such data, the posterior is

$$p(\theta | y) \propto \exp\{[y_2 \log(a\theta_1 + \theta_2) - (1+a)\theta_1 - \theta_2]/\tau\}.$$

As $\tau \rightarrow 0$, $Y_2 \rightarrow 1$ in probability, so the posterior concentrates on the true value of θ $(0, 1)$. Then, since $a\theta_1 + \theta_2 - 1 \rightarrow 0$ as $\theta \rightarrow (0, 1)$, we have

$$p(\theta | y) \propto \exp\left\{\frac{(y_2 - 1)}{\tau}(a\theta_1 + \theta_2 - 1) - \frac{y_2}{2\tau}(a\theta_1 + \theta_2 - 1)^2(1 + o(1)) - \frac{1}{\tau}\theta_1\right\}$$

Therefore, asymptotically, on the event $Y_2 > 0$ (that occurs with high probability, since $Y_2 \xrightarrow{\mathbb{P}} 1$),

$$\frac{a\theta_1 + \theta_2 - 1}{\sqrt{\tau}} | y \sim \mathcal{N}\left(\frac{y_2 - 1}{\sqrt{\tau}y_2}, y_2\right), \quad \frac{\theta_1}{\tau} | y \sim \text{Exp}(1).$$

In practice exponential behaviour of the posterior can be observed for this type of problems (see the histograms of some marginal posteriors of real life SPECT data in Section 6.4).

If we take a product of univariate Jeffreys priors: $p(\theta) = \theta_1^{-1/2}\theta_2^{-1/2}$ for $\theta_j \geq 0$, $j = 1, 2$, that satisfy Assumption P, then the posterior distribution of θ in a neighbourhood of θ^* on the event $Y_2 > 0$ is approximately

$$\frac{a\theta_1 + \theta_2 - 1}{\sqrt{\tau}} | y \sim \mathcal{N}\left(\frac{y_2 - 1}{\sqrt{\tau}y_2}, y_2\right), \quad \frac{\theta_1}{\tau} | y \sim \Gamma(1/2, 1).$$

EXAMPLE 5 (Poisson likelihood 2). Consider $Y_i \sim \text{Pois}(\theta)$, $i = 1, \dots, n$, independently, where the true value is $\theta^* = 0$. In this case, $\mathbb{P}(Y_i = 0) = 1$.

Then, $\tau = 1/n$ and

$$f_Y(\theta) = -\frac{1}{n} \sum_i Y_i \log(\theta) + \theta = \theta$$

with probability 1. Therefore, $f^*(\theta) = \theta$ that achieves the minimum over $[0, \infty)$ at $\theta = 0 = \theta^*$. Consider an improper prior for θ with density $p(\theta) = \theta^{\alpha-1}$ with some $\alpha > 0$. Case $\alpha = 1/2$ corresponds to the Jeffreys prior. All assumptions of Theorem 1 are satisfied for this model.

In this case, the exact posterior distribution for θ is $\Gamma(\alpha, n)$, i.e. $n\theta \mid Y \sim \Gamma(\alpha, 1)$ which agrees with Theorem 1, and the exact 95% credible interval for θ is $\left[0, \frac{\gamma_\alpha(5\%)}{n}\right]$ where $\gamma_\alpha(5\%)$ is the 95% percentile of $\Gamma(\alpha, 1)$ distribution. For $\alpha = 1/2$, the credible interval is $[0, 1.92/n]$, for $\alpha = 0.05$, the credible interval is $[0, 0.27/n]$. By decreasing α to 0, we can construct the credible interval of arbitrarily small length for fixed n , although the adequate coverage probability for any θ^* will be achieved for large n .

If $\theta^* > 0$, then the asymptotic distribution is Gaussian, since the assumptions of the standard Bernstein-von Mises theorem are satisfied.

EXAMPLE 6 (Binomial distribution). Consider a problem of estimating the unknown probability of the Binomial distribution

$$Y_i \sim \text{Bin}(n_i, \theta_i) \quad \text{independently, } i = 1, \dots, p,$$

for $\theta_i \in [0, 1]$, and that some of the true values θ_i^* are 0. We assume that all $\theta_i^* < 1$ (if $\theta_i^* = 1$ for some i , consider $n_i - Y_i$ as data and $1 - \theta_i^*$ as the corresponding parameter). We study the limit of the posterior distribution for large n_i for all $i = 1, \dots, p$ such that $n_i/n \rightarrow \omega_i \in (0, 1)$ where $n = \sum_{i=1}^p n_i$ and p is fixed. This situation is not covered by the standard BvM theorem. Consider a conjugate Beta prior $\theta_i \sim B(\alpha, \varkappa)$ independently, with some fixed $\alpha, \varkappa > 0$. In this case, $\tau = 1/n$,

$$\begin{aligned} f_Y(\theta) &= -\frac{1}{n} \sum_{i=1}^p [Y_i \log(\theta_i) + (n_i - Y_i) \log(1 - \theta_i)], \\ \nabla_i f_Y(\theta) &= -\frac{Y_i}{n\theta_i} + \frac{(n_i - Y_i)}{n(1 - \theta_i)}, \end{aligned}$$

and the limits as $n \rightarrow \infty$ are

$$\begin{aligned} f^*(\theta) &= \lim_{n \rightarrow \infty} f_Y(\theta) = -\sum_{i=1}^p \omega_i [\theta_i^* \log(\theta_i) + (1 - \theta_i^*) \log(1 - \theta_i)], \\ \nabla_i f^*(\theta) &= \lim_{n \rightarrow \infty} \nabla_i f_Y(\theta) = -\omega_i [\theta_i^* / \theta_i - (1 - \theta_i^*) / (1 - \theta_i)], \\ \nabla_{ii} f^*(\theta) &= \lim_{n \rightarrow \infty} \nabla_{ii} f_Y(\theta) = \omega_i [\theta_i^* / \theta_i^2 + (1 - \theta_i^*) / (1 - \theta_i)^2]. \end{aligned}$$

Therefore, Assumption *M* is satisfied (see also Example 2 in Section 3.2). In particular, if $\theta_i^* \neq 0$ ($i \in \bar{S}$), $\nabla_i f^*(\theta^*) = 0$ and $\nabla_{ii} f^*(\theta^*) = \frac{\omega_i}{\theta_i^*(1-\theta_i^*)}$, where the latter coincides with the Fisher information for a simple observation, and if $\theta_i^* = 0$ ($i \in S$), $\nabla_i f^*(\theta^*) = \omega_i$ and $\nabla_{ii} f^*(\theta^*) = \omega_i/(1-\theta_i^*)$. For this model, S^* is always empty. Assumptions *B*, *C*, *S* and *P* are satisfied for the prior distribution with $\alpha, \varkappa > 0$, since $p(\theta) \prod_{j \in S} \theta_j^{1-\alpha}$ is bounded away from 0 and infinity in a neighbourhood of θ^* .

Assumption *L* holds by Lemma 3 with $C_{\delta 0} = C_Y \delta$, $C_{\delta 1} = \min_{i \in S} n_i/n$, where

$$C_Y = \min_{i \in \bar{S}} \left[\frac{Y_i}{n(\theta_i^* + \delta_0/\sqrt{p_0})} + \frac{n_i - Y_i}{n(1 - \theta_i^* + \delta_0/\sqrt{p_0})} - \frac{|Y_i - n_i \theta_i^*|/\sqrt{n}}{\sqrt{n} \delta_0 \theta_i^* (1 - \theta_i^*)} \right]$$

is bounded away from 0 and infinity with high probability for large n , due to inequality $\log(1+x) - x \leq -\frac{a}{a+1}x$ for $x > a > 0$, for large enough n and due to Assumption *S* that $\delta_0 \sqrt{n} \rightarrow \infty$ as $n \rightarrow \infty$.

Therefore, here $\Omega_{00} = \text{diag} \left(\frac{\omega_i}{\theta_i^*(1-\theta_i^*)} i \in \bar{S} \right)$, $a_1 = (\omega_i, i \in S)$, $a_0 = \left(\frac{Y_i - n_i \theta_i^*}{\omega_i \sqrt{n}} \right)_{i \in \bar{S}}$. Applying Theorem 1, the asymptotic posterior distribution for large n_i is approximately

$$(\sqrt{n}(\theta_{\bar{S}} - \theta_{\bar{S}}^*), n(\theta_S - \theta_S^*)) \mid Y \sim \mathcal{N}_{p_0}(a_0, \Omega_{00}^{-1}) \times \Gamma_{p_1}(a_1, \alpha).$$

The corresponding asymptotic $(1-\beta)100\%$ credible interval for θ_i , $i \in \bar{S}$, is

$$\left[\theta_i^* + \frac{Y_i - n_i \theta_i^*}{\omega_i \sqrt{n}} \pm z_{\beta/2} \frac{\sqrt{\theta_i^*(1-\theta_i^*)}}{\sqrt{n \omega_i}} \right]$$

as given by the standard BvM theorem, and for θ_i , $i \in S$ it is $\left[0, \frac{\gamma_{\beta, \alpha_i}}{n \omega_i} \right]$. The $(1-\beta)100\%$ credible interval for θ can be constructed as the product of marginal credible intervals for θ_i with confidence level $(1-\beta/p)100\%$ instead of $(1-\beta)100\%$ (Bonferroni-type correction for multiple testing). The asymptotic $(1-\beta)100\%$ high posterior density region is of the type

$$\mathbb{R}_\beta = \left\{ \theta \in [0, 1]^p : \sum_{i \in \bar{S}} \frac{\omega_i (\theta_i - \theta_i^* - a_{0i}/\sqrt{n})^2}{\theta_i^* (1 - \theta_i^*)} + \sum_{i \in S} [\omega_i \theta_i - (\alpha - 1) \log \theta_i] \leq \frac{C_\beta}{n} \right\}$$

where C_β is chosen in such a way that $\mathbb{P}(\theta \in \mathbb{R}_\beta \mid Y) = 1 - \beta$.

Now we consider a model that is used in variance comparison analysis in mixed effects models under the same type of asymptotics as in [Vu and Zhou \(1997\)](#).

EXAMPLE 7. Consider a mixed effects model

$$Y_{ij} \mid \beta_i \sim \mathcal{N}(\mu + \beta_i, \sigma^2), \quad \beta_i \sim \mathcal{N}(0, \theta)$$

independently, for $i = 1, \dots, n$ and $j = 1, \dots, m$. Here there are n classes with m observations in each, and the parameter of interest here is the contribution of the classes that is characterised by parameter $\theta \in \Theta = [0, \infty)$ where the value $\theta = 0$ corresponds to the absence of the random effects β_i . We study the asymptotic concentration of the posterior distribution when the number of classes n grows while m , the number of observations per class, remains fixed. We consider a prior distribution for θ with density $p(\theta) \propto \theta^{\alpha-1} e^{-b\theta}$ for $\alpha > 0$ and $b \geq 0$ which satisfies Assumption P and which is improper when $b = 0$. Note that the inverse Gamma prior with density $p(\theta) \propto \theta^{-\alpha-1} e^{-b/\theta}$ leads to slow convergence, since it has an essential singularity at 0.

We start with the case of known μ and σ , fixing $\mu = 0$ and $\sigma = 1$ without loss of generality. Integrating out β_i , we have the following marginal likelihood given θ :

$$p(y \mid \theta) = C_y \left[\theta + \frac{1}{m} \right]^{-n/2} \exp \left\{ -\frac{1}{2(\theta + 1/m)} \sum_{i=1}^n \bar{y}_i^2 \right\},$$

where $\bar{y}_i = \frac{1}{m} \sum_{j=1}^m y_{ij}$. The true distribution of data given $\theta = \theta^*$ is $\bar{Y}_i \sim \mathcal{N}(0, \theta^* + 1/m)$, independently. If $\theta^* > 0$, then the model is regular and the standard BuM theorem applies. Now we consider the case $\theta^* = 0$.

In our notation, we have $\tau = 1/n$ and

$$f_Y(\theta) = \frac{1}{2(\theta + 1/m)} \frac{1}{n} \sum_{i=1}^n \bar{Y}_i^2 + \frac{1}{2} \log(\theta + 1/m) \xrightarrow{\mathbb{P}} f^*(\theta)$$

with $f^*(\theta) = \frac{1}{2m(\theta + 1/m)} + \frac{1}{2} \log(\theta + 1/m)$ since $\mathbb{E} \bar{Y}_i^2 = \theta^* + 1/m = 1/m$. The first derivative of $f^*(\theta)$ is

$$\nabla f^*(\theta) = -\frac{1}{2m(\theta + 1/m)^2} + \frac{1}{2(\theta + 1/m)},$$

which implies that the function is minimised at $\theta = 0 = \theta^*$, since the second derivative at $\theta = 0$ is $\nabla^2 f^*(0) = m^2/2 > 0$. Thus, $\Omega_{00} = m^2/2$ and, due to the Central Limit Theorem, the mean is

$$a_0 = -\frac{1}{m\sqrt{n}} \sum_{i=1}^n (m\bar{Y}_i^2 - 1) \sim \mathcal{N}(0, 2/m^2)$$

asymptotically for large n , since $m\bar{Y}_i^2 \sim \chi_1^2$ for all i independently.

The limits of the derivatives of $f_Y(\theta)$ coincide with the corresponding derivatives of $f^*(\theta)$, hence, Assumption M is satisfied for this model. Assumptions B, S and C also hold. Assumption L is satisfied due to Lemma 3 that holds with $C_{\delta_0} = \delta_0 C_{\theta^*, m} \left[1 + \frac{1}{2\delta_0\sqrt{n}} \sum_i \frac{\bar{Y}_i^2 - 1/m - \theta^*}{\sqrt{n}} \right]$ for any $\theta^* \geq 0$ which is bounded with probability close to 1 since $\delta_0\sqrt{n} \rightarrow \infty$ as $n \rightarrow \infty$.

Therefore, by Theorem 1, the approximate posterior distribution of $\sqrt{n}\theta$ has density

$$p_{\theta\sqrt{n}}(x | y) \approx C_{\alpha, m, a_0} x^{\alpha-1} e^{-m^2(x-a_0)^2/4}, \quad x \geq 0.$$

For $\alpha \in (0, 1]$, if $a_0(\omega) > 0$ (asymptotically for a half of possible data sets), this distribution is bimodal, with modes at 0 and at $a_0(\omega)$, and for the other half of possible data sets, the mode is at 0. If $\alpha > 1$, then the distribution is unimodal, with the mode at $\max(0, a_0(\omega))$. Therefore, to improve the recovery of $\theta^* = 0$, α should be in $(0, 1)$ which corresponds to smaller credible interval for θ . In particular, for large n , the MAP is always 0 if $\alpha \in (0, 1)$. Recall that if $\theta^* > 0$ and n is large enough, the standard BvM theorem holds for this prior.

Now we consider the case where the parameters (μ, σ^2, θ) are estimated jointly with a continuous prior for (μ, σ^2) whose density is bounded away from 0 and infinity at the true value (μ^*, σ^*) . Then,

$$\begin{aligned} f_Y(\mu, \sigma^2, \theta) &= \frac{1}{2\sigma^2} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m (Y_{ij} - \bar{Y}_i)^2 + \frac{1}{2\sigma^2(\theta + 1/m)} \frac{1}{n} \sum_{i=1}^n (\bar{Y}_i - \mu)^2 \\ &\quad + \frac{\log(\theta + 1/m)}{2} + \frac{m}{2} \log(\sigma^2), \\ f^*(\mu, \sigma^2, \theta) &= \frac{(m-1)\sigma^{*2}}{2\sigma^2} + \frac{(\mu - \mu^*)^2 + \sigma^{*2}/m}{2\sigma^2(\theta + 1/m)} + \frac{1}{2} \log(\theta + 1/m) + \frac{m}{2} \log(\sigma^2), \end{aligned}$$

since $\mathbb{E}(\bar{Y}_i - \mu)^2 = \frac{\sigma^{*2}}{m} + (\mu - \mu^*)^2$ and $\mathbb{E} \sum_{j=1}^m (Y_{ij} - \bar{Y}_i)^2 = (m-1)\sigma^{*2}$. The function $f^*(\mu, \sigma^2, \theta)$ is minimised at $\mu = \mu^*$, $\sigma = \sigma^*$ and $\theta = \theta^* = 0$, with zero gradient and the matrix of the second order derivatives is

$$\Omega_{00} = \nabla^2 f^*(\mu^*, \sigma^{*2}, \theta^*) = \begin{pmatrix} m/\sigma^{*2} & 0 & 0 \\ 0 & 0.5m/\sigma^{*4} & 0.5m/\sigma^{*2} \\ 0 & 0.5m/\sigma^{*2} & 0.5m^2 \end{pmatrix}.$$

If $\alpha = 1$, then the approximate joint posterior distribution of $\sqrt{n}(\theta - \theta^*, \mu - \mu^*, \sigma^2 - \sigma^{*2})$ is Gaussian truncated to $\theta - \theta^* = \theta \geq 0$ with the bias

as given in Theorem 1 and precision matrix Ω_{00} , or, equivalently, with the covariance matrix

$$\Omega_{00}^{-1} = \begin{pmatrix} \frac{\sigma^{*2}}{m} & 0 & 0 \\ 0 & \frac{2\sigma^{*4}}{m-1} & -\frac{2\sigma^{*2}}{m(m-1)} \\ 0 & -\frac{2\sigma^{*2}}{m(m-1)} & \frac{2}{m(m-1)} \end{pmatrix}.$$

Note that θ and σ^2 are asymptotically correlated, with correlation $-m^{-1/2}$.

If $\alpha \neq 1$, then the approximate posterior has a more complicated form, with an additional factor $\theta^{\alpha-1}$ to the above Gaussian density. As before, the closer α to 0, the smaller the posterior credible interval for θ . However, if m is not very large, improving inference for small θ^* (i.e. choosing small α) will affect inference for σ^2 asymptotically, since they are correlated.

6. Asymptotic behaviour of the posterior distribution for SPECT.

6.1. *Approximation of the posterior distribution.* Consider the SPECT model defined in Section 2, and θ^* with some zero coordinates. Assumptions of Theorem 1 were verified in Examples 1, 3, 4 (Assumptions M, B, C, S, L), and the log cosh Markov random field prior distribution satisfies Assumption P with $\alpha_j = 1$ for all j . Here

$$\nabla f^*(\theta^*) = - \sum_{i: y_i^* \neq 0} y_i^* A_i^T / (A_i \theta^*) + \sum_{i=1}^n A_i = \sum_{i \in Z} A_i^T$$

which is non-zero of $Z = \{i \in \{1, \dots, n\} : y_i^* = 0\}$ is not empty. Hence, nonregularity arises from the elements where there are no detected photons ($y_i^* = 0$) and the likelihood degenerates: $\mathbb{P}_{y_i^*}(Y_i = 0) = 1$ but, since $A_i \neq 0$, it gives us information about those θ_j where $A_{ij} \neq 0$, i.e. on $S = \{j : \theta_j^* = 0 \& \sum_{i \in Z} A_{ij} \neq 0\}$ where the limiting distribution of θ_S / σ^2 is exponential with parameter $a_1 = \sum_{i \in Z} A_{iS}^T$.

Parameter $(\theta_{\bar{S}} - \theta_{\bar{S}}^*) / \sigma$ has approximately truncated Gaussian distribution with parameters

$$\begin{aligned} \Omega_{00} &= A_{\bar{Z}, \bar{S}}^T \text{diag}(1/[y^*]_{\bar{Z}}) A_{\bar{Z}, \bar{S}}, \\ a_0 &= \Omega_{00}^{-1} A_{\bar{Z}, \bar{S}}^T \tilde{Y} / \sigma, \end{aligned}$$

where \tilde{Y} is a vector with coordinates $Y_i / y_i^* - 1$ for $i \in \bar{Z}$. Truncation takes place for parameters θ_{S^*} with $S^* = \{j : \theta_j^* = 0 \& \sum_{i \in Z} A_{ij} = 0\}$.

If the vector of means of the Poisson distribution $y^* = A\theta^*$ has only positive coordinates (Z is empty), this model is regular and the posterior distribution of $(\theta - \theta^*) / \sigma$ is approximately truncated Gaussian.

6.2. *Practical implications of the approximate posterior.* In this section, we briefly discuss some practical implications of Theorem 1. There are well-developed methods for SPECT reconstruction using our model, using Markov chain Monte Carlo computation, delivering not only approximate, simulation-consistent, posterior means, but also variances; see Weir (1997) for a fully Bayesian reconstruction. In this context, the theorem provides valuable knowledge which can enrich the interpretation of numerical results, enabling approximate probabilistic inference.

Inferential questions of real interest, including (a) quantitative inference about amounts of radio-labelled tracer within specified regions of interest, or (b) tests for significance of apparent hot- or cold-spots, can be answered using approximate posterior distributions for *linear combinations* $\lambda^T \theta$ of elements of θ , and are particularly amenable to treatment in this way. More specifically, if for any non-empty set of pixels $R \subseteq \{1, 2, \dots, p\}$, α^R denotes the vector with elements $\alpha_j^R = 1/|R|$ for $j \in R$, 0 otherwise, then to deal with case (a) we can take $\lambda = \alpha^R$ to deliver $\lambda^T \theta$ as the average concentration of tracer in region R , and for case (b) take $\lambda = \alpha^{R_1} - \alpha^{R_2}$ to give the difference in average concentration in region R_1 compared to R_2 .

6.3. *Construction of approximation for the posterior distribution for SPECT model.* To construct an approximation of the posterior distribution, we use estimates of unknown parameters. We use the marginal posterior modes estimate $\hat{\theta}$, $\hat{\theta}_i = \operatorname{argmax} p(\theta_i|y)$, instead of θ^* , $\hat{y} = A\hat{\theta}$ instead of y^* ,

$$\hat{S} = \{j : \nabla_j f_y(\hat{\theta}) > 0\}, \quad \hat{Z} = \{i : \hat{y}_i = 0\}.$$

A more robust way to estimate S would be to use $\hat{S}_\epsilon = \{j : \nabla_j f_y(\hat{\theta}) > \epsilon\}$ for some small enough $\epsilon > 0$, however, a sensitivity to the choice of ϵ would need to be investigated.

Then, the whole approximate posterior of $z = (\theta - \hat{\theta})$ is

$$\phi(z) = \prod_{j \in \hat{S}} [\hat{a}_j / (2\sigma^2)] (2\pi\sigma^2)^{-p_0/2} [\det(\hat{\Omega})]^{1/2} \exp\{-z_{\hat{S}}^T \hat{\Omega} z_{\hat{S}} / (2\sigma^2) - z_{\hat{S}}^T \hat{a} / \sigma^2\},$$

where

$$\begin{aligned} \hat{\Omega} &= \nabla_{\hat{S}, \hat{S}}^2 f_y(\hat{\theta}) = \sum_{i \notin \hat{Z}} y_i / [\hat{y}_i]^2 A_{i, \hat{S}} A_{i, \hat{S}}^T, \\ \hat{a} &= \nabla_{\hat{S}} f_y(\hat{\theta}) = \sum_{i \in \hat{Z}} A_{i, \hat{S}}^T. \end{aligned}$$

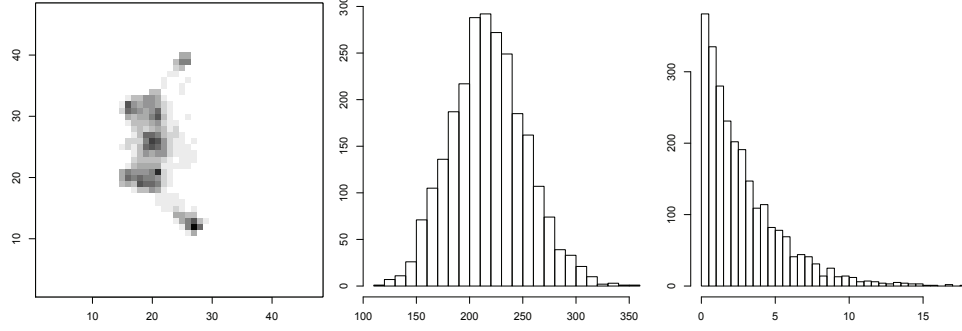


FIG 1. Analysis of real SPECT data: posterior mean reconstruction as a grey-scale image, histogram of marginal posterior for a high-spot pixel (row 12, column 28), and the same for a low-spot pixel (row 12, column 31).

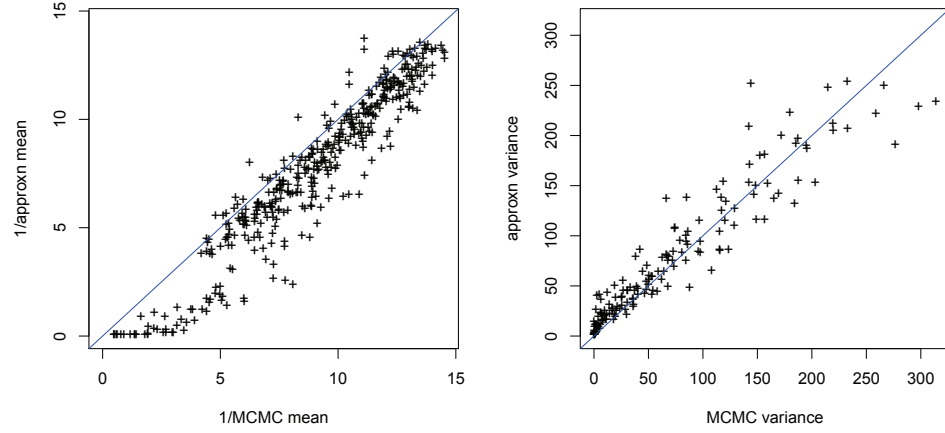


FIG 2. Agreement between (left panel) the elements of $\hat{\mathbf{a}}$ and the reciprocals of the MCMC-computed posterior means of θ , for pixels in \hat{S} , and also that between (right panel) the diagonal elements of $\hat{\Omega}^{-1}$ and the posterior variances of θ for pixels in \hat{S} .

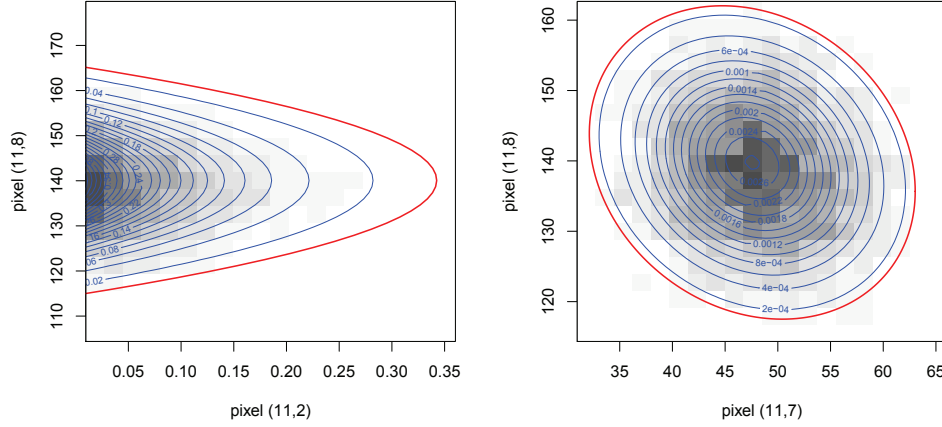


FIG 3. Two bivariate marginals of the posterior, as computed by MCMC (grey-scale image), and the corresponding approximations (contours). In the left panel, one pixel is in \hat{S} and one in \tilde{S} , so the approximation is normal/exponential; in the right panel both pixels are from \tilde{S} , so we have a bivariate normal. The red contour represents the 95% HPD credible region based on the approximation.

6.4. *Finite sample performance.* Finally, we briefly discuss the extent to which the approximation in Theorem 1 holds true for data on the scale of a real SPECT study. A formal assessment of this would entail a major study beyond the scope of this paper, so instead we present selected results from analysis of two data sets based on a SPECT scan of the pelvic region of a human subject.

In the first experiment, the matrix A was constructed according to the model in Green (1990) and Weir (1997), capturing geometry, attenuation and radioactive decay for a setup consisting of 64 projections from a 2-dimensional slice through the patient, each projection yielding an array of 52 photon counts, corresponding to a spatial resolution of 0.57cm. The data set was obtained from Bristol Royal Infirmary; the total photon count was 45652; individual counts ranged from 0 to 85, averaging 13.7. Reconstruction was performed on a 48×48 square grid, with pixel size 0.64cm, using the log cosh prior with hyperparameters fixed at $\gamma = 25$ and $\delta = 8$, was obtained using a simple MCMC sampler. We employed 20000 sweeps of a deterministic-raster-scan single-pixel random walk Metropolis sampler on a square-root scale for θ , chosen to avoid extremes in acceptance rate at high- and low-spots in the image.

Figure 1 shows selected aspects of this analysis; see caption for details. Our tentative conclusion from this is that the marginal posterior distributions for individual pixels x_j do appear to be approximately normal in high-spots and approximately exponential in low-spots, consistent with the theoretical limits presented in Theorem 1.

A second experiment was focussed on a more precise and quantitative assessment of the approximation to the posterior derived in the previous section. The setup is the same as in the first experiment, except at half the resolution, so that reconstruction was on a 24×24 grid, with pixel size 1.28cm. The corresponding A matrix is now better-conditioned, and p is only 576, so that manipulation of the matrices is entirely tractable. Synthetic data was generated using this A and a ‘ground truth’ obtained from an approximate MAP reconstruction from the same real data set as used above, yielding a total photon count of 138310, and individual counts ranging up to 243. 50000 sweeps of the MCMC sampler were used, and the prior settings were $\gamma = 200$, $\delta = 8$.

Figure 2 displays the agreement between the elements of \hat{a} and the reciprocals of the MCMC-computed posterior means of θ , for pixels in \hat{S} , and also that between the diagonal elements of $\hat{\Omega}^{-1}$ and the posterior variances of θ for pixels in $\hat{\tilde{S}}$.

Figure 3 displays two bivariate marginals of the posterior, as computed by MCMC, and the corresponding approximations. In the left panel, one component is in \hat{S} and one in $\hat{\tilde{S}}$, so the approximation is normal/exponential; in the right panel both components are from $\hat{\tilde{S}}$, so we have a bivariate normal.

We conclude that for this realistic/modest-scale SPECT reconstruction problem, the small-variance asymptotics of this paper provide a good approximation to the posterior even for $\sigma^2 = 1$.

7. Discussion. When the posterior distribution concentrates on the boundary, we have showed that the classic Bernstein–von Mises theorem, stating the limit of the posterior distribution recentred and rescaled by \sqrt{n} for n independent random variables, does not hold. Instead, the limit differs in two respects, in directions towards the boundary: the limiting distribution is a Gamma distribution, and the appropriate scale is n , i.e. the convergence is faster. The shape of the limiting Gamma distribution depends on the behaviour of the prior distribution on the boundary. Parallel to the boundary, however, the classic version of Bernstein–von Mises theorem is applicable. Our results also extend the Bernstein–von Mises theorem to the case of non-iid observations.

A remarkable property of Bayesian inference in the model considered here is that on the boundary, we can recover the unknown parameter with arbitrarily high precision, that is, deliver a credible interval that is arbitrarily small when the true value of the parameter is 0, without adversely affecting inference when the parameter is non-zero. (Of course, a relatively smaller value for the noise parameter σ^2 will be needed to preserve asymptotic efficiency of Bayesian inference for the case the true value of the parameter is non-zero and close to the boundary, with prior density changing fast in its neighbourhood.)

In this case, the limit of the posterior does not depend on the data, in contrast to what is found in other nonregular problems, such as those considered by Ghosal and Samanta (1995), Ghosh *et al.* (1994) and Chernozhukov and Hong (2004). In their case, the density of the errors has a jump whose location depends on the unknown parameter which also leads to a different posterior distribution with a faster convergence rate whose location is shifted by a random variable that does depend on data.

The nonasymptotic version of the main result shows that other parameters of the model can also affect convergence in practice, such as the smallest eigenvalues of the precision matrices in the Gaussian part of the limit and the smallest parameter of the scale of the Gamma distributions.

An interesting direction for future work is to study both the behaviour of the posterior distribution, and the question of optimal prior specification, in a framework where the spatial resolution is infinitely refined, placing smoothness class constraints on θ^* .

References.

- Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion). *J. Roy. Statist. Soc. B*, **48**, 259–302.
- Bochkina, N. (2012). Consistency of the posterior distribution in generalised linear inverse problems. Inverse Problems, under revision.
- Chernozhukov, V. and Hong, H. (2004). Likelihood estimation and inference in a class of nonregular econometric models. *Econometrica*, **72**, (5), 1445–80.
- Douc, R., Moulines, E., Olsson, J., and van Handel, R. (2011). Consistency of the maximum likelihood estimator for general hidden Markov models. *The Annals of Statistics*, **39**, (1), 474–513.
- Dudley, R. M. and Haughton, D. (2002). Asymptotic normality with small relative errors of posterior probabilities of half-spaces. *The Annals of Statistics*, **30**, (5), 1311–44.

- Erkanli, A. (1994). Laplace approximations for posterior expectations when the mode occurs at the boundary of the parameter space. *Journal of the American Statistical Association*, **89**, (425), 250–8.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, (425), 721–41.
- Ghosal, S., Ghosh, J. K., and Samanta, T. (1995). On convergence of posterior distributions. *Annals of Statistics*, **23**, 2145–52.
- Ghosal, S. and Samanta, T. (1995). Asymptotic behaviour of Bayes estimates and posterior distributions in multiparameter nonregular cases. *Math. Methods Statist.*, **4**, 361–88.
- Ghosh, J. K., Ghosal, S., and Samanta, T. (1994). Stability and convergence of posterior in non-regular problems. In *Statistical Decision Theory and Related Topics*, (ed. S. S. Gupta and J. O. Berger), pp. 183–99. Springer.
- Green, P. J. (1990). Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Transactions on Medical Imaging*, **9**, 84–93.
- Hirano, K. and Porter, J. R. (2003). Asymptotic efficiency in parametric structural models with parameter-dependent support. *Econometrica*, **71**, (5), 1307–38.
- Ibragimov, I. A. and Has'minskij, R. Z. (1981). *Statistical estimation: asymptotic theory*. Springer, New York.
- Johnstone, I. M. and Silverman, B. W. (1990). Speed of estimation in positron emission tomography and related inverse problems. *The Annals of Statistics*, **18**, 251–80.
- Kleijn, B. J. K. and van der Vaart, A. W. (2012). The Bernstein-von Mises theorem under misspecification. *Electronic Journal of Statistics*, **6**, 354–81.
- Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *University of California Publications in Statistics*, **1**, 277–330.
- Le Cam, L. and Yang, G. (1990). *Asymptotics in statistics: some basic concepts*. Springer, New York.
- Moran, P. A. P. (1971). Maximum-likelihood estimation in non-standard conditions. *Mathematical Proceedings of the Cambridge Philosophical Society*, **70**, (3), 441–50.

- Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, **135**, (3), 370–84.
- Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, (398), 605–10.
- Shapiro, A. (2000). On the asymptotics of constrained local M-estimators. *The Annals of Statistics*, **28**, (3), 948–60.
- Spokoiny, V. (2012). Parametric estimation. Finite sample theory. *The Annals of Statistics*, **40**, 2877–909.
- van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.
- Vu, H. T. V. and Zhou, S. (1997). Generalization of likelihood ratio tests under nonstandard conditions. *The Annals of Statistics*, **25**, (2), 897–916.
- Weir, I. S. (1997). Fully Bayesian reconstructions from single-photon emission computed tomography data. *Journal of the American Statistical Association*, **92**, (437), 49–60.

APPENDIX A: PROOFS

A.1. Proof of boundary-related properties.

PROOF OF LEMMA 1. 1. Condition: the gradient $\nabla f^*(\theta^*)$ should not be in the polar cone at θ^* , i.e. $\nabla_j f^*(\theta^*) \geq 0$ for $j: \theta_j^* = 0$ and $\nabla_j f^*(\theta^*) = 0$ for $j: \theta_j^* > 0$. Therefore, for $j \in S$ - where $\nabla_j f^*(\theta^*) > 0$, $\theta_j^* = 0$, and for $j \in \bar{S}$, $\nabla_j f^*(\theta^*) = 0$.

We also have a set $S^* \subseteq \bar{S}$ where $\theta_j^* = 0$ and $\nabla_j f^*(\theta^*) = 0$.

2. Consider function $\tilde{f}^*(x) = f^*(\tilde{\theta})$ where $\tilde{\theta}_{\bar{S}} = x$ and $\tilde{\theta}_S = \theta_S^*$, i.e. restricted to the set \bar{S} of coordinates. This function has a unique global minimum at $x = \theta_{\bar{S}}^*$, and $\nabla \tilde{f}^*(x) = 0$. Function $\tilde{f}^*(x)$ is twice differentiable with respect to x with a continuous derivative in a neighbourhood of $\theta_{\bar{S}}^*$. Hence, we must have $\nabla^2 \tilde{f}^*(\theta_{\bar{S}}^*)$ is a positive semi-definite matrix. \square

PROOF OF LEMMA 2. Since $\exists c_0, c_1 > 0: \{v \in B_2(0, c_0) : v_{S^*} \geq 0\} \times [0, c_1]^{p_1} \subseteq V(\Theta - \theta^*)$, we have, as $\sigma \rightarrow 0$,

$$D^{-1}V(\Theta - \theta^*) \supseteq \{v \in B_2(0, c_0/\sigma) : v_{S^*} \geq 0\} \times [0, c_1/\sigma^2]^{p_1} \rightarrow \mathbb{R}^{p_0 - p_0^*} \times \mathbb{R}_+^{p_0^* + p_1},$$

$$\text{i.e. } \mathcal{V}^* = \mathbb{R}^{p_0 - p_0^*} \times \mathbb{R}_+^{p_0^* + p_1}$$

For $\delta_0 \leq c_0$ and $\delta_1 \leq c_1$ and $\delta_0/\sigma \rightarrow \infty$ and $\delta_1/\sigma^2 \rightarrow \infty$, as $\sigma \rightarrow 0$, the sets

$$D^{-1}V(B_\delta(\theta^*) - \theta^*) = \{v \in B_2(0, \delta_0/\sigma) : v_{S^*} \geq 0\} \times [0, \delta_1/\sigma^2]^{p_1}$$

monotonically increase to \mathcal{V}^* which implies the statement of the lemma. \square

A.2. Proof of the main result. We start with a series of lemmas.

LEMMA 4. *Let set $\Theta \subseteq [0, \infty)^p$ be such that its only boundary points are θ with at least one zero coordinate: $\partial\Theta = \cup_{j=1}^p \{\theta \in \Theta : \theta_j = 0\}$.*

Assume function $F(\theta) : \Theta \rightarrow \mathbb{R}$ satisfies the following conditions with some $\theta^ \in \Theta$ and some $S \subset \{1, \dots, p\}$ such that $\theta_S^* = 0$:*

1. $\exists \nabla F(\theta)$ for $\theta \in B_\delta(\theta^*)$ for some $\delta = (\delta_0, \delta_1)$, where $B_\delta(\theta^*)$ is defined by (8).
2. $\exists \Delta_{F,1}, \Delta_{F,2} > 0$ such that $\Delta_{F,1} \rightarrow 0$ and $\Delta_{F,2} \rightarrow 0$ as $\delta_0 \rightarrow 0$ and $\delta_1 \rightarrow 0$ such that for any $\theta \in B_\delta(\theta^*)$,

$$\max_{j \in S} |\nabla_j F(\theta) - \nabla_j F(\theta^*)| \leq \Delta_{F,1}, \quad \max_{i,j \in \bar{S}} |\nabla_{ij} F(\theta) - \nabla_{ij} F(\theta^*)| \leq \Delta_{F,2}.$$

Then, for any $\theta \in B_\delta(\theta^*)$,

$$\begin{aligned} F(\theta) - F(\theta^*) &\leq (\theta_S - \theta_S^*)^T [\nabla_S F(\theta^*) + \Delta_{F,1} 1_{|S|}] + (\theta_{\bar{S}} - \theta_{\bar{S}}^*)^T \nabla_{\bar{S}} F(\theta^*) \\ &\quad + (\theta_{\bar{S}} - \theta_{\bar{S}}^*)^T [\nabla_{\bar{S},\bar{S}} F(\theta^*) + \Delta_{F,2} I_{|\bar{S}|}] (\theta_{\bar{S}} - \theta_{\bar{S}}^*) / 2, \end{aligned}$$

$$\begin{aligned} F(\theta) - F(\theta^*) &\geq (\theta_S - \theta_S^*)^T [\nabla_S F(\theta^*) - \Delta_{F,1} 1_{|S|}] + (\theta_{\bar{S}} - \theta_{\bar{S}}^*)^T \nabla_{\bar{S}} F(\theta^*) \\ &\quad + (\theta_{\bar{S}} - \theta_{\bar{S}}^*)^T [\nabla_{\bar{S},\bar{S}} F(\theta^*) - \Delta_{F,2} I_{|\bar{S}|}] (\theta_{\bar{S}} - \theta_{\bar{S}}^*) / 2. \end{aligned}$$

The proof easily follows from the Taylor expansion of $F(\theta)$ at θ^* up to the first order term for $j \in S$ and up to the second order term for $j \in \bar{S}$, and bounding the corresponding first and second order derivatives using the assumptions of the lemma.

Denote $\text{int}(\Theta)$ the interior of set Θ .

LEMMA 5. Suppose function $F(\theta)$ satisfies conditions of Lemma 4 with set $S \cup S^*$, and the following conditions hold for a σ -finite measure $\pi(d\theta)$ on Θ .

1. For $\theta \in B_\delta(\theta^*)$, $\exists p(\theta) \geq 0$: $\pi(d\theta) = p(\theta)d\theta$
2. $\exists \Delta_\pi \in (0, 1)$, $C_\pi \in (0, \infty)$ and $\exists \alpha_j \in (0, \infty)$ for $j \in S \cup S^*$ such that for $\theta \in B_\delta(\theta^*)$,

$$C_\pi(1 - \Delta_\pi) \leq p(\theta) \times \prod_{j \in S \cup S^*} \theta_j^{1-\alpha_j} \leq C_\pi(1 + \Delta_\pi).$$

Then, for any $\mathcal{B} \subset B_\delta(\theta^*)$, as $\tau \rightarrow 0$,

$$\begin{aligned} &\int_{\mathcal{B}} \exp\{-F(\theta)/\tau\} \pi(d\theta) \leq C_\pi \exp\{-F(\theta^*)/\tau\} (1 + \Delta_\pi) \\ &\times \int_{\mathcal{B}-\theta^*} \mu(dv; \tau^{-1}[\nabla_S F(\theta^*) + \Delta_{F,1} 1_{|S|}], \alpha, -\nabla_{\bar{S}} F(\theta^*)/\tau, \tau^{-1}[\nabla_{\bar{S},\bar{S}} F(\theta^*) + \Delta_{F,2} I_{|\bar{S}|}]), \\ &\int_{\mathcal{B}} \exp\{-F(\theta)/\tau\} \pi(d\theta) \geq C_\pi \exp\{-F(\theta^*)/\tau\} (1 - \Delta_\pi) \\ &\times \int_{\mathcal{B}-\theta^*} \mu(dv; \tau^{-1}[\nabla_S F(\theta^*) - \Delta_{F,1} 1_{|S|}], \alpha, -\nabla_{\bar{S}} F(\theta^*)/\tau, \tau^{-1}[\nabla_{\bar{S},\bar{S}} F(\theta^*) - \Delta_{F,2} I_{|\bar{S}|}]), \end{aligned}$$

where measure $\mu(dv; a_1, \alpha, a_0, \Sigma)$ for $a_1, \alpha \in (0, \infty)^{p_1}$, $a_0 \in \mathbb{R}^{p_0}$ and a $p_0 \times p_0$ positive definite matrix Σ is defined by

$$(14) \quad \frac{\mu(dv; a_1, \alpha, a_0, \Sigma)}{dv} = \prod_{j=1}^{p_0^*} v_{0,p_0-p_0^*+j}^{\alpha_j-1} \prod_{i=1}^{p_1} v_{1,i}^{\alpha_i-1} e^{-a_1^T v_1 - v_0^T \Sigma v_0 / 2 + v_0^T a_0},$$

for $v = (v_0^T, v_1^T)^T \in \mathbb{R}^{p_0} \times (0, \infty)^{p_1}$.

LEMMA 6. Consider function $f_Y(\theta)$ defined in Section 3.1 and assume that Assumptions B, S and C hold. Define the following events for some $\delta_{*0}, \delta_{*1} \rightarrow 0$ as $\sigma \rightarrow 0$ (that exist due to Assumptions M, C and S):

$$(15) \quad \begin{aligned} \mathcal{A}_0 &= \{\omega : \|\nabla_S f_{Y(\omega)}(\theta^*) - a_1\|_\infty \leq \delta_{*1}\}, \\ \mathcal{A}_1 &= \{\omega : \|\nabla_{\bar{S}, \bar{S}}^2 f_{Y(\omega)}(\theta^*) - \Omega_{00}\| \leq \delta_{*0}\}, \\ \mathcal{A}_2 &= \{\omega : \|\nabla_S f_{Y(\omega)}(\theta^*) - \nabla_S f_{Y(\omega)}(\theta)\|_\infty \leq \delta_{*1} \forall \theta \in B_{\delta}(\theta^*)\}, \\ \mathcal{A}_3 &= \{\omega : \|\nabla_{\bar{S}, \bar{S}}^2 f_{Y(\omega)}(\theta^*) - \nabla_{\bar{S}, \bar{S}}^2 f_{Y(\omega)}(\theta)\| \leq \delta_{*0} \forall \theta \in B_{\delta}(\theta^*)\}. \end{aligned}$$

Then, on event $\mathcal{A} = \mathcal{A}_0 \cap \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$, for $\theta \in B_{\delta}(\theta^*)$,

$$\begin{aligned} f_Y(\theta) - f_Y(\theta^*) &\leq \tilde{a}^T(\theta_S - \theta_S^*) + (\theta_{\bar{S}} - \theta_{\bar{S}}^*)^T \tilde{\Omega}_{00}(\theta_{\bar{S}} - \theta_{\bar{S}}^*)/2 + (\theta_{\bar{S}} - \theta_{\bar{S}}^*)^T \nabla f_Y(\theta^*), \\ f_Y(\theta) - f_Y(\theta^*) &\geq \bar{a}^T(\theta_S - \theta_S^*) + (\theta_{\bar{S}} - \theta_{\bar{S}}^*)^T \bar{\Omega}_{00}(\theta_{\bar{S}} - \theta_{\bar{S}}^*)/2 + (\theta_{\bar{S}} - \theta_{\bar{S}}^*)^T \nabla f_Y(\theta^*), \end{aligned}$$

where

$$\begin{aligned} \tilde{\Omega}_{00} &= \Omega_{00} + 2\delta_{*0}I_{p_0}, & \tilde{a} &= a_1 + 2\delta_{*1}\mathbf{1}, \\ \bar{\Omega}_{00} &= \Omega_{00} - 2\delta_{*0}I_{p_0}, & \bar{a} &= a_1 - 2\delta_{*1}\mathbf{1}. \end{aligned}$$

PROOF. Applying Lemma 4 with $F(\theta) = f_{Y(\omega)}(\theta)$ for $\omega \in \mathcal{A}_0 \cap \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$, we have the following upper bound:

$$\begin{aligned} f_Y(\theta) - f_Y(\theta^*) &\leq (\theta_S - \theta_S^*)^T [\nabla_S f_Y(\theta) + \delta_{*1}] + (\theta_{\bar{S}} - \theta_{\bar{S}}^*)^T \nabla_{\bar{S}} f_Y(\theta) \\ &\quad + (\theta_{\bar{S}} - \theta_{\bar{S}}^*)^T [\nabla_{\bar{S}, \bar{S}}^2 f_Y(\theta) + \delta_{*0}I_{|\bar{S}|}](\theta_{\bar{S}} - \theta_{\bar{S}}^*)/2 \\ &\leq (\theta_S - \theta_S^*)^T [a_1 + 2\delta_{*1}\mathbf{1}_{|S|}] + (\theta_{\bar{S}} - \theta_{\bar{S}}^*)^T \nabla_{\bar{S}} f_Y(\theta) \\ &\quad + (\theta_{\bar{S}} - \theta_{\bar{S}}^*)^T [\Omega_{00} + 2\delta_{*0}I_{|\bar{S}|}](\theta_{\bar{S}} - \theta_{\bar{S}}^*)/2 \end{aligned}$$

since $\theta_S - \theta_S^* = \theta_S$ is a vector with non-negative coordinates, which is the first statement of the lemma. Applying the remaining inequalities on events \mathcal{A}_k , we obtain the second statement of the lemma. \square

PROOF OF THEOREM 1. Consider a neighbourhood of θ^* , $B_{\delta}(\theta^*) = (\theta^* + B_{\delta}) \cap \Theta$, where $B_{\delta} = B_{2,p_0}(0, \delta_0) \times B_{\infty,p_1}(0, \delta_1)$. Denote $v = (v_0^T, v_1^T)^T$ where $v_0 = (\theta_{\bar{S}} - \theta_{\bar{S}}^*)/\sigma$ and $v_1 = (\theta_S - \theta_S^*)/\sigma^2$, with the Jacobian of this change of variables being $\sigma^{p_0+2p_1}$. Here we use $\alpha = (\alpha_0, \alpha_1)$ (i.e. after the permutation of the coordinates).

For the rescaled parameter v , consider the corresponding neighbourhood B_R with $R_0 = \delta_0/\sigma$, $R_1 = \delta_1/\sigma^2$:

$$B_R = B_2(0, R_0) \times [0, R_1]^{p_1} \cap D_{\sigma}^{-1}V(\Theta - \theta^*),$$

Under Assumption S, $\delta_k \rightarrow 0$ and $R_k \rightarrow \infty$, $k = 0, 1$. These conditions hold, for instance, with $R_k = C_k[-\log \sigma]^{a_k}$ for some positive constants C_k and a_k . Hence, by Lemma 2, set B_R becomes \mathcal{V}^* as $\sigma \rightarrow 0$.

For small enough δ_0, δ_1 , $B_R = [B_{2,p_0}(0, R_0)\mathbf{1}_{\mathcal{V}_0}] \times [0, R_1]^{p_1}$. This condition is satisfied if $\|\theta_S^*\| \geq \delta_0$ and $\delta_k \leq c_k$ where constants c_k are given in Assumption B.

The triangle inequality for the total variation norm gives us

$$\begin{aligned} \|\mathbb{P}_{S(\theta-\theta^*)|Y} - \mu^*\|_{TV} &\leq \|\mathbb{P}_{S(\theta-\theta^*)|Y} \mathbf{1}_{B_R} - \mu^* \mathbf{1}_{B_R}\|_{TV} \\ (16) \quad &+ \|\mu^* \mathbf{1}_{B_R} - \mu^*\|_{TV} + \|\mathbb{P}_{S(\theta-\theta^*)|Y} \mathbf{1}_{B_R} - \mathbb{P}_{S(\theta-\theta^*)|Y}\|_{TV}, \end{aligned}$$

where the balls B_R are defined above. Here $\mu \mathbf{1}_{B_R}$ is a probability measure μ truncated to B_R and normalised to be a probability measure.

If measures μ_1, μ_2 are absolutely continuous with respect to some measure μ with densities f and g respectively, then the total variation norm can also be written as

$$\|\mu_1 - \mu_2\|_{TV} = 2 \int_{\Theta} (f - g)_+ d\mu,$$

where $(x)_+ = \max(x, 0)$ (van der Vaart 1998). In each of the summands in the upper bound (16), the first measure is absolutely continuous with respect to the second one, so we will use this expression to evaluate the total variation norm.

We start with the distance between the truncations of the rescaled posterior distribution and the limit on B_R . By Lemmas 5 and 6 (whose assumptions are satisfied due to Assumptions B, C and S), on event $\mathcal{A} = \mathcal{A}_0 \cap \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$ defined in Lemma 6, for any $\mathcal{B} \subseteq B_\delta(\theta^*)$, with $\mathcal{B}_v = D_\sigma^{-1}V(\mathcal{B} - \theta^*) \subseteq B_R$, we have

$$\begin{aligned} &\int_{\mathcal{B}} \exp\{-[f_y(\theta) - f_y(\theta^*)]/\sigma^2\} \pi(d\theta) \geq \sigma^{p_0 - p_0^* + \sum_{j=1}^{p_0^*} \alpha_{0,j} + 2 \sum_{j=1}^{p_1} \alpha_{1,j}} C_\pi (1 + \Delta_\pi) \\ &\times \int_{\mathcal{B}_v} \exp\{-\tilde{a}^T v_1\} \prod_i v_i^{\alpha_i} \exp\left\{-\|\tilde{\Omega}_{00}^{1/2} v_0\|^2/2 - v_0^T \nabla f_y(\theta^*)/\sigma\right\} dv \\ &= J_\sigma C_\pi (1 + \Delta_\pi) \mu(\mathcal{B}_v; \tilde{a}, \alpha, -\nabla_{\tilde{S}} f_y(\theta^*)/\sigma, \tilde{\Omega}_{00}), \end{aligned}$$

where $\prod_i v_i^{\alpha_i}$ here and below stands for $\prod_{j=1}^{p_0^*} v_{0,p_0-p_0^*+j}^{\alpha_{0,j}-1} \prod_{j=1}^{p_1} v_{1,j}^{\alpha_{1,j}-1}$, measure $\mu(dv; a_1, \alpha, a_0, \Sigma)$ is defined by (14) and $J_\sigma = \sigma^{p_0 - p_0^* + \sum_{j=1}^{p_0^*} \alpha_{0,j} + 2 \sum_{j=1}^{p_1} \alpha_{1,j}}$.

Similarly, using Lemmas 5 and 6, we obtain an upper bound on event \mathcal{A} :

$$\begin{aligned} & \int_{\mathcal{B}} \exp \{ -[f_y(\theta) - f_y(\theta^*)]/\sigma^2 \} \pi(d\theta) \leq J_\sigma C_\pi (1 - \Delta_\pi) \\ & \times \int_{\mathcal{B}_v} \prod_i v_i^{\alpha_i - 1} \exp \{ -\bar{a}^T v_1 \} \exp \left\{ -\|\bar{\Omega}_{00}^{1/2} v_0\|^2/2 - v_0^T \nabla f_y(\theta^*)/\sigma \right\} dv \\ & = J_\sigma C_\pi (1 - \Delta_\pi) \mu(\mathcal{B}_v; \bar{a}, \alpha, -\nabla_{\bar{S}} f_y(\theta^*)/\sigma, \bar{\Omega}_{00}). \end{aligned}$$

To simplify the notation, denote

$$\bar{\mu}(dv) = \mu(dv; \bar{a}, \alpha, -\nabla_{\bar{S}} f_y(\theta^*)/\sigma, \bar{\Omega}_{00}), \quad \tilde{\mu}(dv) = \mu(dv; \tilde{a}, \alpha, -\nabla_{\bar{S}} f_y(\theta^*)/\sigma, \tilde{\Omega}_{00}).$$

These measures are finite if matrices $\tilde{\Omega}_{00}$, $\bar{\Omega}_{00}$ are positive definite, $\nabla_{\bar{S}} f_y(\theta^*)/\sigma$ is finite with probability $\rightarrow 1$ as $\sigma \rightarrow 0$, and all components of vectors \tilde{a} , \bar{a} are positive.

Measure $\tilde{\mu}$ is finite since $\nabla_{\bar{S}} f_y(\theta^*)/\sigma$ is finite with high probability due to Assumption C, and all other parameters are positive or positive definite. Measure $\bar{\mu}$ is finite if $2\delta_{*1} < \min_j a_{1,j}$ and $2\delta_{*0} < \lambda_{\min}(\Omega_{00})$ which hold for small enough δ_{*0}, δ_{*1} .

For $\mathcal{V}^* = \mathbb{R}^{p_0 - p_0^*} \times \mathbb{R}_+^{p_0^* + p_1}$ and $\mathcal{B}_v = \mathcal{B}_1 \times B_\infty(0, r_1)$, we have

$$\begin{aligned} \mu(\mathcal{V}^*; a, \alpha, a_0, \Sigma) &= \prod_{i=1}^{p_1} [a_i^{1-\alpha_i} \Gamma(\alpha_i)] \Phi_{p_0^*}(\mathcal{V}_0; \Sigma^{-1} a_0, \Sigma, \alpha_0), \\ \mu(\mathcal{B}_v; a, \alpha, a_0, \Sigma) &= \mu(\mathcal{V}^*; a, \alpha, a_0, \Sigma) \Phi_{p_0^*}(\mathcal{B}_1; \Sigma^{-1} a_0, \Sigma, \alpha_0) \prod_{j=1}^{p_1} \Gamma(r_1; \alpha_{1,j}, a_{1,j}), \end{aligned}$$

where probability measure $\Phi_{p_0^*}(\cdot; a_0, \Omega_{00}, \alpha)$ is defined by (7), $\Gamma(r; \alpha, a)$ is the cumulative distribution function of distribution $\Gamma(\alpha, a)$ and $\alpha = (\alpha_0, \alpha_1)$.

Hence, the posterior density of $\mathcal{S}(\theta - \theta^*)$ normalised by the posterior measure of B_R is bounded on \mathcal{A} by

$$\frac{1 - \Delta_\pi}{1 + \Delta_\pi} \frac{\tilde{\mu}(dv)}{\bar{\mu}(B_R)} \leq \frac{d p(\mathcal{S}(\theta - \theta^*) | Y)}{p(B_R | Y)} \leq \frac{\bar{\mu}(dv)}{\tilde{\mu}(B_R)} \frac{1 + \Delta_\pi}{1 - \Delta_\pi}.$$

Therefore, the total variation distance between the rescaled posterior distribution and its limit, both truncated to B_R , is bounded on \mathcal{A} by

$$\begin{aligned} \|\mathbb{P}_{\mathcal{S}(\theta - \theta^*)|Y} \mathbf{1}_{B_R} - \mu^* \mathbf{1}_{B_R}\|_{TV} &\leq 2 \int_{B_R} \left[\frac{\mathbb{P}(dv | Y) \mu^*(B_R)}{\mathbb{P}(B_R | Y) \mu^*(dv)} - 1 \right]_+ \frac{\mu^*(dv)}{\mu^*(B_R)} \\ &\leq 2 \int_{B_R} \left[\frac{\bar{\mu}(dv)}{\tilde{\mu}(B_R)} \frac{\mu^*(B_R)}{\mu^*(dv)} \frac{(1 + \Delta_\pi)^2}{(1 - \Delta_\pi)^2} - 1 \right]_+ \frac{\mu^*(dv)}{\mu^*(B_R)} \end{aligned}$$

Now, $\frac{\mu^*(dv)}{\mu^*(B_R)} = \frac{\mu_0(dv)}{\mu_0(B_R)}$ where $\mu_0(dv) = \mu(dv; a_1, \alpha, \Omega_{00}a_0, \Omega_{00})$. Then,

$$\frac{\bar{\mu}(dv)}{\mu_0(dv)} = \exp\{2\delta_{*1}\mathbf{1}^T v_1 + 2\delta_{*0}\|v_0\|^2/2\},$$

which implies, with $a_0 = -\Omega_{00}^{-1}\nabla_{\bar{S}}f_Y(\theta^*)/\sigma$,

$$\begin{aligned} & \frac{\bar{\mu}(dv)}{\mu_0(dv)} \frac{\mu_0(B_R)}{\tilde{\mu}(B_R)} = \exp\{2\delta_{*0}\|v_0\|^2/2 + 2\delta_{*1}\mathbf{1}^T v_1\} \\ & \times \frac{\int_{B_R} \prod_i v_i^{\alpha_i-1} \exp\{-a_1^T v_1\} \exp\{-\|\Omega_{00}^{1/2} v_0\|^2/2 + v_0^T \Omega_{00} a_0\} dv}{\int_{B_R} \prod_i v_i^{\alpha_i-1} \exp\{-(a_1 + 2\delta_{*1}\mathbf{1})^T v_1 - \|\tilde{\Omega}_{00}^{1/2} v_0\|^2/2 + v_0^T \Omega_{00} a_0\} dv}. \end{aligned}$$

To show that this expression is greater than 1, it is sufficient to show that for any $\mathcal{B} \subseteq \{v_0 : (v_0^T, v_1^T)^T \in B_R\}$, the following expression is positive:

$$\begin{aligned} & \int_{\mathcal{B}} e^{-\|\Omega_{00}^{1/2} w\|^2/2 + w^T \Omega_{00} a_0} dw - \int_{\mathcal{B}} e^{-\|\tilde{\Omega}_{00}^{1/2} w\|^2/2 + w^T \Omega_{00} a_0} dw \\ & = \int_{\mathcal{B}} \prod_i w_i^{\alpha_i-1} e^{-\|\tilde{\Omega}_{00}^{1/2} w\|^2/2 + w^T \Omega_{00} a_0} [\exp\{\delta_{*0}\|w\|^2\} - 1] dw > 0 \end{aligned}$$

which is indeed the case. Therefore, on \mathcal{A} , $\frac{\bar{\mu}(dv)}{\mu_0(dv)} \frac{\mu_0(B_R)}{\tilde{\mu}(B_R)} \geq 1$ and hence

$$\begin{aligned} & \|\mathbb{P}_{S(\theta-\theta^*)|Y} \mathbf{1}_{B_R} - \mu^* \mathbf{1}_{B_R}\|_{TV} \leq 2 \int_{B_R} \left[\frac{\bar{\mu}(dv)}{\mu(B_R)} \frac{\mu^*(B_R)}{\mu^*(dv)} \frac{(1+\Delta_\pi)^2}{(1-\Delta_\pi)^2} - 1 \right] \frac{\mu^*(dv)}{\mu^*(B_R)} \\ & = 2 \left[\frac{\bar{\mu}(B_R)}{\mu(B_R)} \frac{(1+\Delta_\pi)^2}{(1-\Delta_\pi)^2} - 1 \right] = 2 \frac{\bar{\mu}(B_R) - \tilde{\mu}(B_R)}{\tilde{\mu}(B_R)} \frac{(1+\Delta_\pi)^2}{(1-\Delta_\pi)^2} + 2 \left[\frac{(1+\Delta_\pi)^2}{(1-\Delta_\pi)^2} - 1 \right]. \end{aligned}$$

The difference of measures $\bar{\mu}(B_R) - \tilde{\mu}(B_R)$ is bounded by

$$\begin{aligned} & \int_{B_R} \prod_i w_i^{\alpha_i-1} e^{-w_0^T \tilde{\Omega}_{00} w_0/2 + w_0^T \Omega_{00} a_0 - \tilde{a} w_1} \left[e^{\delta_{*0}\|w_0\|^2 + 2\delta_{*1}\mathbf{1}_{p_1}^T w_1} - 1 \right] dw \\ & \leq \int_{B_R} \prod_i w_i^{\alpha_i-1} [\delta_{*0}\|w_0\|^2 + 2\delta_{*1}\mathbf{1}_{p_1}^T w_1] e^{-w_0^T \tilde{\Omega}_{00} w_0/2 + w_0^T \Omega_{00} a_0 - \tilde{a} w_1} dw \\ & \leq \left[2\delta_{*0}E_\Phi + 2\delta_{*1} \sum_{j=1}^{p_1} (\alpha_{1,j}/\bar{a}_j) \right] \bar{\mu}(\mathcal{V}^*) \end{aligned}$$

due to inequality $e^x - 1 \leq xe^x$ for $x > 0$, and with E_Φ defined by

$$E_\Phi = \int_{\mathcal{V}_0} \|w\|^2 \Phi_{p_0^*}(dw; -\bar{\Omega}_{00}^{-1}\nabla_{\bar{S}}f_Y(\theta^*), \bar{\Omega}_{00}, \alpha_0),$$

which is finite. Therefore,

$$\begin{aligned} \|\mathbb{P}_{S(\theta-\theta^*)|Y} \mathbf{1}_{B_R} - \mu^* \mathbf{1}_{B_R}\|_{TV} &\leq 2 \left[\frac{(1+\Delta_\pi)^2}{(1-\Delta_\pi)^2} - 1 \right] \\ &+ 2 \frac{\bar{\mu}(\mathcal{V}^*)}{\bar{\mu}(B_R)} \frac{(1+\Delta_\pi)^2}{(1-\Delta_\pi)^2} \left[2\delta_{*0} E_\Phi + 2\delta_{*1} \sum_{j=1}^{p_1} (\alpha_{1,j}/\bar{a}_j) \right], \end{aligned}$$

which goes to zero since $\delta_{*k} \rightarrow 0$ and $\Delta_\pi \rightarrow 0$ as $\sigma \rightarrow 0$.

For $R_0, R_1 \rightarrow \infty$,

$$\frac{\bar{\mu}(B_R)}{\bar{\mu}(\mathcal{V})} = \Phi_{p_0^*}(B_2(0, R_0); -\tilde{\Omega}_{00}^{-1} \nabla_{\bar{S}} f_Y(\theta^*)/\sigma, \tilde{\Omega}_{00}, \alpha_0) \times \prod_{j=1}^{p_1} \Gamma(R_1; \alpha_{1,j}, \tilde{a}_j),$$

which is close to 1 for large R_0 and R_1 . Therefore, $\|\mathbb{P}_{S(\theta-\theta^*)|Y} \mathbf{1}_{B_R} - \mathbb{P}_{S(\theta-\theta^*)|Y} \mathbf{1}_{B_R}\|_{TV} \rightarrow 0$ as $\sigma \rightarrow 0$.

The total variation distance between the limit measure and its truncation to B_R is bounded by

$$\|\mu^* - \mu^* \mathbf{1}_{B_R}\|_{TV} \leq 2\mu^*(\bar{B}_R)$$

as $\sigma \rightarrow 0$, since $R_0, R_1 \rightarrow \infty$.

The total variation distance between the posterior distribution and its truncation to B_R is bounded by

$$\begin{aligned} \|\mathbb{P}_{(S(\theta-\theta^*)|Y)} \mathbf{1}_{B_R} - \mathbb{P}_{(S(\theta-\theta^*)|Y)}\|_{TV} &\leq 2\mathbb{P}_{(S(\theta-\theta^*)|Y)}(\bar{B}_R) \\ &= 2 \frac{\int_{\Theta \setminus B_\delta(\theta^*)} \exp\{-(f_y(\theta) - f_y(\theta^*))/\sigma^2\} d\pi(x)}{\int_{\Theta} \exp\{-(f_y(\theta) - f_y(\theta^*))/\sigma^2\} d\pi(x)} \\ &\leq 2[C_\pi(1+\Delta_\pi)\bar{\mu}(B_R)]^{-1} \Delta_0(\delta), \end{aligned}$$

where $\Delta_0(\delta)$ defined by (11) is

$$\Delta_0(\delta) = \sigma^{p_0-p_0^*+\sum_{j=1}^{p_0^*} \alpha_{0,j}+2\sum_{j=1}^{p_1} \alpha_{1,j}} \int_{\Theta \setminus B_\delta(\theta^*)} \exp\{-(f_y(\theta) - f_y(\theta^*))/\sigma^2\} \pi(d\theta).$$

By Assumption L, with probability $\rightarrow 0$, $\Delta_0(\delta) \rightarrow 0$ as $\sigma \rightarrow 0$, and $\bar{\mu}(B_R) \rightarrow \mu_0(B_R) > 0$.

Combining these bounds, we have that on \mathcal{A} ,

$$\begin{aligned} \|\mathbb{P}_{S(\theta-\theta^*)|Y} - \mu^*\|_{TV} &\leq 2\mu^*(\bar{B}_R) + 2[C_\pi(1+\Delta_\pi)\bar{\mu}(B_R)]^{-1} \Delta_0(\delta) \\ &+ 4 \frac{\bar{\mu}(\mathcal{V}^*)}{\bar{\mu}(B_R)} \frac{(1+\Delta_\pi)^2}{(1-\Delta_\pi)^2} \left[\delta_{*0} E_\Phi + \delta_{*1} \sum_{j=1}^{p_1} (\alpha_{1,j}/\bar{a}_j) \right] + \frac{8\Delta_\pi}{(1-\Delta_\pi)^2} \rightarrow 0 \end{aligned}$$

and $\mathbb{P}(\mathcal{A}) \rightarrow 1$ as $\sigma \rightarrow 0$, which gives the statement of the theorem. \square

PROOF OF PROPOSITION 1. In the proof of Theorem 1, we derived that on \mathcal{A} with $\delta_{*k} = M_k \delta_k$, $\rho = M_2 \sigma$,

$$\|\mathbb{P}_{S(\theta-\theta^*)|Y} - \mu^*\|_{TV} \leq 2\mu^*(\overline{B_R}) + 2C_\Delta \Delta_0(\delta) + 2C_0 \delta_0 + 2C_1 \delta_1 + C_2 \Delta_\pi$$

where

$$\begin{aligned} C_0 &= M_0 C_A E_\Phi, & C_1 &= M_1 C_A \sum_{j=1}^{p_1} \alpha_{1,j} / \bar{a}_j, \\ C_2 &= \frac{8}{(1-\Delta_\pi)^2}, & C_\Delta &= [C_\pi(1+\Delta_\pi)\tilde{\mu}(B_R)]^{-1}, \end{aligned}$$

and

$$\begin{aligned} C_A &= 2 \frac{\bar{\mu}(\mathcal{V}^*)}{\tilde{\mu}(B_R)} \frac{(1+\Delta_\pi)^2}{(1-\Delta_\pi)^2} \\ &= 2 \frac{\Phi_{p_0^*}(\mathcal{V}_0; -\tilde{\Omega}_{00}^{-1} \nabla_{\bar{S}} f_Y(\theta^*)/\sigma, \tilde{\Omega}_{00}, \alpha_0)}{\Phi_{p_0^*}(\mathcal{V}_0; -\bar{\Omega}_{00}^{-1} \nabla_{\bar{S}} f_Y(\theta^*)/\sigma, \bar{\Omega}_{00}, \alpha_0)} \times \prod_{j=1}^{p_1} \left[\frac{a_{1,j} + 2\delta_{*1}}{a_{1,j} - 2\delta_{*1}} \right]^{\alpha_{1,j}} \frac{(1+\Delta_\pi)^2}{(1-\Delta_\pi)^2}. \end{aligned}$$

Here $E_\Phi = \int_{\mathcal{V}_0} \|w\|^2 \Phi_{p_0^*}(dw; -\bar{\Omega}_{00}^{-1} \nabla_{\bar{S}} f_Y(\theta^*), \bar{\Omega}_{00}, \alpha_0)$. If $\alpha_j = 1 \forall j \in S^*$,

$$E_\Phi = [|\bar{\Omega}_{00}^{-1} \nabla_{\bar{S}} f_Y(\theta^*)/\sigma|^2 + \text{trace}(\bar{\Omega}_{00}^{-1})] \frac{(2\pi)^{p_0/2}}{[\det(\bar{\Omega}_{00})]^{1/2}} e^{\|\bar{\Omega}_{00}^{-1/2} \nabla_{\bar{S}} f_Y(\theta^*)/\sigma\|^2/2}.$$

Consider the term $\mu^*(\overline{B_R})$:

$$\begin{aligned} \mu^*(\overline{B_R}) &= 1 - \int_{v_0 \in B_2(0, \delta_0/\sigma): v_0^* \geq 0} \mu_0^*(dv_0) \int_{v_1 \in B_\infty(0, \delta_1/\sigma^2)} \mu_1^*(dv_1) \\ &\leq 1 - \int_{v_0 \in B_2(0, \delta_0/\sigma): v_0^* \geq 0} \mu_0^*(dv_0) + \int_{v_1 \notin B_\infty(0, \delta_1/\sigma^2)} \mu_1^*(dv_1) \\ &= \mu_0^*(\overline{B_{2+}(0, \delta_0/\sigma)}) + \mu_1^*(\overline{B_\infty(0, \delta_1/\sigma^2)}) \end{aligned}$$

using inequality $1 - xy \leq 1 - x + 1 - y$ for $x, y \in (0, 1)$. We can also use

$$\mu_1^*(\overline{B_\infty(0, \delta_1/\sigma^2)}) \leq p_1 [1 - \min_j \Gamma(\delta_1/\sigma^2; \alpha_{1,j}, a_{1,j})],$$

and, changing to polar coordinates and denoting $p_{\alpha 0} = p_0 + \sum_{j=1}^{p_0^*} (\alpha_j - 1)$,

$$\begin{aligned} \mu_0^* \left(\overline{B_{2+}(0, \delta_0/\sigma)} \right) &\leq C_\Phi \int_{\delta_0/\sigma - \|a_0\|}^{\infty} r^{p_{\alpha 0}-1} e^{-\lambda_{\min}(\Omega_{00})r^2/2} dr \\ &\quad \times \int_{\sum_{i=1}^{p_0} w_i^2=1} \prod_{j=1}^{p_0^*} w_j^{\alpha_{0j}-1} dw \\ &\leq C_{\alpha 0} \left[1 - \Gamma((\delta_0/\sigma - \|a_0\|)^2/2; p_{\alpha 0}/2, \lambda_{\min}(\Omega_{00})) \right], \end{aligned}$$

where $C_{\alpha 0} = C_\Phi V_{p_0} 2^{-p_0} D(p_0, p_0^*, \alpha_0)$, C_Φ is the normalising constant, V_{p_0} is the surface area of the unit sphere in p_0 dimensions, $D(p_0, p_0^*, \alpha_0)$ is the normalising constant for p_0 -dimensional Dirichlet distribution with parameter $(1, \dots, 1, \alpha_{0,1}, \dots, \alpha_{0,p_0^*})$.

Collecting the non-asymptotic conditions on δ_k in the proof of Theorem 1, we have

$$\delta_1 < a_{\min}/(2M_1), \quad \delta_0 < \lambda_{\min}(\Omega_{00})/(2M_0), \quad \delta_0 \leq \|\theta_S^*\|, \quad \delta_0 \leq c_0, \quad \delta_1 \leq c_1.$$

Thus, we have the required statement. \square

A.3. Auxiliary results.

PROOF OF LEMMA 3. For small enough σ , under Assumption B and S on δ_0, δ_1 ,

$$B_\delta(\theta^*) \supset \theta^* + (-\delta_0/\sqrt{p_0}, \delta_0/\sqrt{p_0})^{p_0-p_0^*} \times [0, \delta_0/\sqrt{p_0}]^{p_0^*} \times [0, \delta_1/\sigma^2]^{p_1},$$

due to $\|x\|_\infty \leq r$ implying $\|x\|_2 \leq r\sqrt{p_0}$ for $x \in \mathbb{R}^{p_0}$.

Under the assumptions of the lemma, for small enough σ ,

$$\begin{aligned}
 & \int_{\Theta \setminus B_\delta(\theta^*)} e^{-(f_y(\theta) - f_y(\theta^*)) / \sigma^2} \pi(d\theta) \\
 \leq & C_{\pi 0}(\delta) \sum_{j \in \bar{S} \setminus S^*} \int_0^{\theta_j^* - \delta_0 / \sqrt{p_0}} \theta_j^{\alpha_j - 1} e^{-C_{\delta 0} |\theta_j - \theta_j^*| / \sigma^2} d\theta_j \\
 + & C_{\pi 0}(\delta) \sum_{j \in \bar{S}} \int_{\delta_0 / \sqrt{p_0}}^\infty e^{-C_{\delta 0} v_j / \sigma^2} dv_j + C_{\pi 0}(\delta) \sum_{j \in S} \int_{\delta_1}^\infty e^{-C_{\delta 1} v_j / \sigma^2} dv_j \\
 \leq & C_{\pi 0}(\delta) \sum_{j \in \bar{S} \setminus S^*} \sigma^{\alpha_j} e^{-C_{\delta 0}(\theta_j^* - \sigma) / \sigma^2} \\
 + & C_{\pi 0}(\delta) \sum_{j \in \bar{S} \setminus S^*} \left[\sigma^{\alpha_j - 1} I(\alpha_j < 1) + \theta_j^{\alpha_j - 1} I(\alpha_j \geq 1) \right] \frac{\sigma^2}{C_{\delta 0}} e^{-C_{\delta 0} \delta_0 / (\sqrt{p_0} \sigma^2)} \\
 + & p_0 C_{\pi 0}(\delta) e^{-C_{\delta 0} \delta_0 / [\sqrt{p_0} \sigma^2]} \sigma^2 / C_{\delta 0} + p_1 C_{\pi 0}(\delta) e^{-C_{\delta 1} \delta_1 / \sigma^2} \sigma^2 / C_{\delta 1} \\
 \leq & C [C_{\pi 0}(\delta) \sigma^{\min_j(\alpha_j)} + C_{\pi 0}(\delta) \sigma] e^{-C_{\delta 0} \delta_0 / [\sqrt{p_0} \sigma^2]} + p_1 C_{\pi 0}(\delta) e^{-C_{\delta 1} \delta_1 / \sigma^2} \sigma^2 / C_{\delta 1}
 \end{aligned}$$

for a constant C , which implies that

$$\begin{aligned}
 \Delta_0(\delta) &= \sigma^{-\sum_{j=1}^{p_0} \alpha_{0,j} - 2 \sum_{j=1}^{p_1} \alpha_{1,j}} \int_{\Theta \setminus B_\delta(\theta^*)} e^{-(f_y(\theta) - f_y(\theta^*)) / \sigma^2} \pi(d\theta) \\
 &\leq C \sigma^{-\sum_{j=1}^{p_0} \alpha_{0,j} - 2 \sum_{j=1}^{p_1} \alpha_{1,j}} \\
 &\quad \times \left\{ [C_{\pi 0}(\delta) \sigma^{\min_j(\alpha_j)} + C_{\pi 0}(\delta) \sigma] e^{-C_{\delta 0} \delta_0 / [\sqrt{p_0} \sigma^2]} \right. \\
 &\quad \left. + p_1 C_{\pi 0}(\delta) e^{-C_{\delta 1} \delta_1 / \sigma^2} \sigma^2 / C_{\delta 1} \right\} \rightarrow 0
 \end{aligned}$$

as $\sigma \rightarrow 0$ under the assumptions of the theorem. \square

E-MAIL: N.Bochkina@ed.ac.uk

E-MAIL: P.J.Green@bristol.ac.uk

SCHOOL OF MATHEMATICS, UNIVERSITY OF EDINBURGH, EDINBURGH EH9 3JZ, UK. E-MAIL: N.Bochkina@ed.ac.uk

SCHOOL OF MATHEMATICS, UNIVERSITY OF BRISTOL, BRISTOL BS8 1TW, UK, AND

SCHOOL OF MATHEMATICAL SCIENCES, UNIVERSITY OF TECHNOLOGY, SYDNEY, AUSTRALIA. E-MAIL: P.J.Green@bristol.ac.uk